*European Journal of Risk Regulation* (2026), 1–16
doi:10.1017/err.2026.10152

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

# Risk for AI Agents

Federico L.G. Faroldi

CERNAI, Università di Pavia, Italy
Email: federico.faroldi@unipv.it

## Abstract

This paper investigates whether the concept of risk applies to AI agents. The EU AI Act, in force since 2024, grounds its regulatory architecture in a product-based model of risk: one that presupposes an intended purpose, (a bounded space of) reasonably foreseeable misuses, and a tractable distribution of potential harms. This paper argues that these three presuppositions fail structurally for sufficiently general agentic systems. Agents whose behavioural space is open-ended by design cannot be assigned a purpose specific enough to anchor risk identification; their misuse space is generatively open, including novel sequences discoverable by the agent itself; and the probability and severity of harm are subject to modification through the agent's own learning and action on the environment. What fails, however, is the product-style operationalisation of risk (per system, *ex ante*, anchored in intended purpose), rather than risk as such. Given this diagnosis, the paper develops two paths for conceptualising agentic risk: a conservative path, which retains the standard definition of risk and reconstructs it formally for narrow agents (either as expected deviation from an ideal trajectory, or as the probability-weighted sum of constraint violations); and a governance path, which draws an analogy to the normative systems we use to manage biological agents, arguing that sufficiently general AI agents require something closer to law-following than to product-safety assessment. The two paths are compared along five dimensions and shown to be complementary: the first suited to operational risk monitoring within bounded deployments and the second to the governance architecture that determines which agents may legitimately be deployed at all.

**Keywords:** agentic risk; law-following AI; EU AI Act

## I. Introduction and motivation

The most prominent and comprehensive legislation on AI is the EU AI Act, in force since August 2024 in the European Union. The EU AI Act is risk-based: it sorts AI practices and systems into categories of unacceptable risk (prohibited outright under Article 5), high risk (Articles 6–16), limited risk (subject to the transparency obligations of Article 50) and minimal risk, and it imposes correspondingly different requirements on providers. For instance, providers of high-risk systems have to adopt a risk-management system. The EU AI Act, moreover, characterises AI systems as products. Most of such systems, in fact, are supposed to have an intended purpose, and the risk analysis takes it into account together with reasonably foreseeable misuses. The Act does recognise, however, that not all AI systems fit this mould: a separate regime for general-purpose AI models (GPAI) introduces

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use and/or adaptation of the article.

https://doi.org/10.1017/err.2026.10152

the notion of systemic risk (Article 3(65)), and the GPAI provisions do not require an intended purpose in the same sense as the high-risk framework.[1]

Since at least 2024, however, the industry, the media and academia have hailed (or forewarned about) the era of AI agents: systems that can not only act in the world, but have some autonomy in choosing their means, if not their ends. Such agents are often *general*. First, they do not have a specific, limited intended purpose; second, they can be misused in creative, unforeseeable ways by the users; and third, they can often come up with strategies of their own to cause harm.[2]

This paper deals with the question of whether risk, and therefore a risk-based system, can apply to AI agents, and if so, how.

The paper is organised as follows: in Section II, I introduce the concepts of "agent" and "risk" as currently used in the relevant domains; in Section III, I investigate whether "risk" applies to AI agents, arguing that the standard, product-style operationalisation of risk breaks down with sufficiently sophisticated agents. In Section IV, I develop two paths for conceptualising agentic risk: a conservative, formal path for narrow agents and a governance path for general agents. Section V compares the two paths along five dimensions, and Section VI concludes.

## II. Agency and risk: definitions

Before we can explore whether the concept of risk applies to AI agents, we need to clarify what we mean by "agent," "AI agent," and "risk," taking into account current literature and specialised uses in standards and legislation that are relevant for the context of this paper.

### *1. Agents*

#### *a. On agency in general*

The debate on agency has a long history. Agency can be used in a wide sense, whereby everything that causally intervenes on something else can be said to act on it, or in a narrow sense, whereby agency requires intentional action (or other specific requirements: initiation by the agent, second-order desires, etc.). The latter is rooted in the history of Western philosophy (Aristotle, Hume, Anscombe, Davidson, ...), with important differences and variations that are not germane to the current inquiry.

More generally, outside of philosophy, (at least some scholars in) computer science and evolutionary biology (perhaps not so surprisingly) agree: an agent is "a feedback control system in which the individual directs its actions toward goals and controls those actions via perceptually informed decisions."[3,4]

This is a very minimalistic definition: famously, for instance, it applies to thermostats. Philosophers and legal scholars are usually more demanding, requiring either substantive or functional characteristics. In the context of the present work, we can stick to the textbook, minimalistic definition, so as not to exclude too many potential (AI) agents.

[1] See Arts. 51–5 for the systemic-risk obligations applicable to high-impact GPAI models. The structural difference between the GPAI regime and the high-risk regime is discussed further in Section III.

[2] See, e.g., T Korbak and others, "How to Evaluate Control Measures for LLM Agents? A Trajectory from Today to Superintelligence" (arXiv, 2025) <https://arxiv.org/abs/2504.05259>.

[3] M Tomasello, *The Evolution of Agency: Behavioral Organization from Lizards to Humans* (MIT Press 2022) p 19.

[4] See the enormously influential S Russell and P Norvig, *Artificial Intelligence: A Modern Approach* (Pearson 2021), for a version of this understanding of agency.

https://doi.org/10.1017/err.2026.10152

*b. On AI agents*

But what about artificial agents? Both companies and the press have hailed the rise of AI "agents," or "agentic" systems (mostly personal assistants based on large language models), and scholars have been criticising cutting-edge legislation for not being suited to deal with agential systems. But are such systems really agentic in one or more of the senses mentioned above? Kapoor and others[5] maintain that AI systems can be considered "agentic" if either

- they can pursue difficult goals without being instructed in complex environments;
- they can be instructed in natural language and act autonomously without supervision;
- they are systems that are able to use tools, such as web search or programming, or are capable of planning . . .

This is a functional perspective, since it does not really focus on intrinsic characteristics or techniques that such systems need to be "truly" agentic.

In any case, the (i) pursuit of goals, (ii) autonomy of actions and (iii) planning capabilities are all abilities that can be seen as falling out of having intentions and acting on them: more directly when it comes to (i) and (iii), for instance if one considers Bratman's belief–desire–intention (BDI) model,[6] and perhaps less directly related to intentions when it comes to (ii).

There's recently been some philosophical work trying to characterise AI agents in their contemporary form, rather than in their historical form.[7]

Kasirzadeh and Gabriel[8] propose to evaluate AI agency along the following dimensions: autonomy, efficacy, goal complexity and generality.

Dung[9] also proposes a multidimensional account of agency, along the following lines:

- goal-directedness: a being pursues aims;
- autonomy: a being is autonomous to the extent that it can function on its own, without intentional intervention by an outside observer;
- efficacy: a being is efficacious if its goal-directed behaviour affects the world;
- planning: the planning dimension is "the degree to which the algorithmic system is [able] to make decisions that are temporally dependent upon one another to achieve a goal and/or make predictions over a long time horizon";
- intentionality: One is an intentional agent in this sense if and only if one is able to act for reasons.[10]

The International AI Safety Report (IAISR) of January 2025 proposes the following working definition for *AI agents*:[11]

[5] S Kapoor and others, "AI Agents That Matter" (arXiv, 1 July 2024) <https://arxiv.org/abs/2407.01502>.

[6] M E Bratman, *Intention, Plans, and Practical Reason* (Harvard University Press 1987).

[7] L Dung, "Understanding AI Agency" (2025) Philosophical Studies; A Kasirzadeh and I Gabriel, "Characterizing AI Agents for Alignment and Governance" (arXiv, 2025) <https://arxiv.org/pdf/2504.21848>. Cf M Wooldridge, *Reasoning about Rational Agents* (MIT Press 2000); M Wooldridge, *An Introduction to MultiAgent Systems* (2nd edn, Wiley 2009).

[8] Kasirzadeh and Gabriel, *supra*, note 7.

[9] Dung, *supra*, note 7.

[10] For an argument that AI agents should be designed as reason-responsive, see F L G Faroldi, "Reasons-Based Artificial Agents" (2026) 6 AI and Ethics 77.

[11] Y. Bengio, S. Mindermann, D. Privitera, et al. *"International AI Safety Report" (DSIT 2025/001, 2025,* available at <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025> (last accessed 11 August 2026).

https://doi.org/10.1017/err.2026.10152

> "A general-purpose AI which can make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes along the way, and interact with its environment – for example by creating files, taking actions on the web, or delegating tasks to other agents – with little to no human oversight."

The IAISR definition builds generality into the *definition* of an agent rather than treating it as a dimension along which agents vary. Instead, we decouple agency from generality. For the purposes of this paper, we are going to adopt the following definition:

> An *AI agent* is a system that can make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes, and interact with its environment with some degree of autonomy – whether or not it is general-purpose.

This is consistent with the standard understanding in the AI literature.[12] Generality is then treated as a separate, scalar dimension of agency – roughly, the breadth of the task and environment space the system can operate competently within – that runs from narrowly task-specific systems at one end to systems whose behavioural space is as broad as whatever is computable at the other.

Such a definition is, broadly speaking, functionalist and architecture-agnostic.

Moreover, it is quite minimal, in comparison to the characterisations found in the recent literature. However, those dimensions will turn out to matter directly for the question of whether and how risk applies. The argument of Section III will show that the standard notion of risk breaks down even for agents satisfying only this weak definition, provided they score high enough on the generality dimension; and the two paths I will develop in Section IV are, in effect, different strategies for coping with different regions of the autonomy–generality space.

## 2. *Risk*

According to the Stanford Encyclopaedia of Philosophy, "risk"[13] can take five main meanings: (i) an unwanted event; (ii) the cause of an unwanted event; (iii) the probability of an unwanted event; (iv) the expected value of an unwanted event; and (v) the fact that a decision is made under conditions of known probabilities (rather than uncertainty).[14]

Standard organisations such as ISO, and legislators who follow their lead, employ a definition that at first glance appears different: risk as the *combination of the probability of harm and the severity of that harm*. This is an operationalisation of meaning (iv): the expected value of an unwanted event just is, in the simplest case, the product of its probability and its magnitude (severity), which is a way of combining them. The ISO formulation allows these two components to be estimated and compared – often through qualitative instruments such as risk matrices, which discretise the continuous probability-times-severity product into manageable categories.[15]

[12] Russell and Norvig, *supra*, note 4.

[13] S O Hansson, "Risk" in E N Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Summer 2023 edn) <https://plato.stanford.edu/archives/sum2023/entries/risk/>.

[14] For fuller treatments of the concept of risk and its philosophical foundations, see S O Hansson, *The Ethics of Risk: Ethical Analysis in an Uncertain World* (Palgrave Macmillan 2013); N Möller, "The Concepts of Risk and Safety" in S Roeser and others (eds), *Handbook of Risk Theory* (Springer 2012) p 55.

[15] The discretisation introduces well-known distortions: risk matrices can rank a high-probability/low-severity risk above a low-probability/catastrophic one, depending on how cells are scored, even when expected values would reverse the ordering. For a critique, see, e.g., L A Cox, 'What's Wrong with Risk Matrices?' (2008) 28 Risk Analysis 497. For present purposes, however, these distortions are orthogonal to our argument. The expected-value framework requires a cardinal utility scale and a well-defined probability distribution over outcomes,

https://doi.org/10.1017/err.2026.10152

The operationalised notion is the notion adopted by the EU AI Act, namely in Article 3, which is the one we adopt here without further ado.[16]

## III. Does the concept of risk apply to (AI) agents?

With working definitions of "agent" and "risk" in hand, we can now move to ask the central question of this paper: does the concept of risk apply to AI agents? There are already some examples in the literature that are variously labelled under "agentic risk" or "agentic misalignment."[17] Before proceeding, let's distinguish four claims that are often run together in this vicinity, and that are better kept separate. The first is *conceptual*: the product-style operationalisation of risk, as probability-times-severity of harm assessed *ex ante* for a single system against its intended purpose, is inapplicable to sufficiently general agents, because the conditions that make the operationalisation applicable do not obtain for such systems. The second is *doctrinal*: the specific risk-management architecture of the EU AI Act, built around provider obligations and *ex ante* assessment, fails to accommodate agentic systems for the same structural reasons. The third is *practical*: even setting aside conceptual and doctrinal questions, *ex ante* risk assessment is computationally and epistemically infeasible for agents whose behavioural space expands with deployment. The fourth is *normative*: given these failures, a governance approach is probably required. This section argues for the first three claims in that order. The fourth is not established here; it requires the comparative argument of Sections V and IV.2, and I defer it accordingly.

Back to the main question: does risk apply to AI Agents?

Since we focused on the notion of "risk" often employed in product safety legislation or standards, one side of the question can then be seen, for certain purposes, as asking whether agents are products: are the fundamental characteristics of products the same in agents, where such fundamental characteristics are those that, for our purpose, ground the application of the notion of risk to products. Another side of the question can be seen as asking whether the fundamental characteristics of risk apply to agents.[18]

whereas the ISO formulation is deliberately agnostic about these foundations, allowing ordinal severity judgements and qualitative likelihood estimates. This difference in presupposition will matter when we ask, in Section III, whether either framework can be applied to general agents – and the answer will be the same for both, for the same structural reasons.

[16] There is another notion of risk introduced in the EU AI Act, that of systemic risk (Art. 3(65)). I discuss it in main text below, in the context of the GPAI regime.

[17] One concrete example may be Anthropic's hypothesised "agentic misalignment" category (A Lynch and others, "Agentic Misalignment: How LLMs Could Be Insider Threats" (arXiv, 2025) <https://arxiv.org/abs/2510.05179>), which seems to consist of deliberately choosing harmful action without external pressure to do so, in contingent scenarios: "We distinguish agentic misalignment from several other forms of misalignment. First, it differs from exogenous misalignment caused by bad actors prior to model deployment (such as sleeper agents inserted during training). Second, it differs from hypothesized endogenous forms of misalignment like inherent hostility toward humanity, since agentic misalignment is contextually contingent rather than fundamental. Third, while some of the scenarios we study (like corporate espionage) are similar to jailbreaking or prompt injection, others (like blackmail) involve models choosing harmful action without external inducement. This research highlights how models' intentional actions can be a source of harm even without adversarial prompting (ib.)."

[18] Another clarification is in order before proceeding. The critique developed in this section is directed at a specific layer of the EU AI Act: the high-risk architecture of Arts. 6–16, which imposes *ex ante* conformity assessment and risk-management obligations on providers, which constitutes the Act's primary operative mechanism. The GPAI regime (Arts. 51–5) is fairly distinct, and I address it separately below; the point is not that the Act ignores general-purpose systems, but that even its GPAI provisions inherit the conceptual difficulties I am about to describe. The Act's risk-based approach has attracted a substantial doctrinal literature: see M E Kaminski, "Regulating the Risks of AI" (2023) 103 Boston University Law Review 1347, on the structure of risk regulation applied to AI; J Laux, S Wachter and B Mittelstadt, "Trustworthy Artificial Intelligence and the European Union AI Act: On the Conflation of Trustworthiness and Acceptability of Risk" (2024) 18 Regulation & Governance 3, on the

https://doi.org/10.1017/err.2026.10152

Traditional risk requires three fundamental elements: (i) intended purpose, which is a defined, stable function that constrains the possibility space of behaviours, and thus allows risk identification; (ii) foreseeable misuse as a bounded deviation, which can be anticipated (e.g., consider the EU AI Act "reasonably foreseeable misuse" clause) and finally (iii) a combination of known probabilities and severities of harms: in a sense, this requires a closed-world assumption whereby one must know what harms are possible, in order to be able to estimate their probability and severity.

AI agents, if sufficiently general, break these three components. Let's see how. First, *intended purpose*. An intended purpose must be specific enough to constrain behaviour, but AI agents are designed for generality – their value comes from doing whatever users need. We could think about using "help users" as an intended purpose of such systems, but given its breadth, it cannot be used to ground risk identification in a satisfactory way. On the other hand, any purpose statement specific enough to anchor risk assessment mischaracterises the system. Thus, true generality does not merely make intended purpose complex or imprecise but it seems to void the entire category. The EU AI Act seems to acknowledge this implicitly: Articles 6–16, governing high-risk systems, are built around intended purpose as a structural requirement, whereas the GPAI provisions of Articles 51–5 ignore it – possibly a legislative recognition that general-purpose systems resist the intended-purpose framing. Does the GPAI regime, then, escape the critique? I do not think it does. The GPAI provisions introduce the notion of systemic risk (Article 3(65)), defined as a risk "specific to the high-impact capabilities of general-purpose AI models" that can be "propagated at scale across the value chain." But systemic risk remains conceptually continuous with the standard product-risk framework: it still turns on "reasonably foreseeable negative effects," it still presupposes a tractable harm space that can be assessed *ex ante*, and it still places the evaluative burden on the provider's risk-identification capacity. Since systemic risk is still risk, and it's characterised by its footprint, rather than different essential properties, I do not think it changes the analysis significantly in the current setting.[19]

Second, *foreseeable misuse*. Agents have compositional capability: they can chain actions in arbitrary sequences, so that the number of possible action sequences grows considerably. But this is not (just) a computational problem; it's conceptual, since novelty (also) emerges from composition: many harmful sequences involve combinations never seen in training. Consider a coffee machine: we can enumerate misuses because the physical constraint space is small. Compare with a coding agent: the space of possible behaviours approaches whatever is computable. Thus, foreseeability requires the ability to characterise the space of concerns in advance, but for agents, the misuse space is not just large but open-ended and includes novel combinations discoverable by the agent itself.[20]

Third, we come to the *combination of probability and severity* of harms. In the case of agents, new ways to cause harm can be discovered through exploration and reasoning;

conflation of trustworthiness and acceptability of risk; C Novelli and others, "Taking AI Risks Seriously: A New Assessment Model for the AI Act" (2024) 39 AI & Society 2493, and T Mahler, "Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal" in L Colonna and R Greenstein (eds), *Nordic Yearbook of Law and Informatics 2020–2021: Law in the Era of Artificial Intelligence* (Swedish Law and Informatics Research Institute 2022) p 247, on the Act's risk-assessment model and its proportionality logic; and N A Smuha, "Beyond the Individual: Governing AI's Societal Harm" (2021) 10(3) Internet Policy Review 1, on societal harms that escape the individual-risk frame.

[19] For more systematic considerations on systemic risk and agential systems, see F L G Faroldi, "Agentic and Systemic AI Risk" (2025) 23 Revista Iustitia.

[20] Empirical support for this claim comes from recent work on LLM-based agents demonstrating that such systems can autonomously generate attack strategies and harmful action sequences not present in their training data (Lynch and others, *supra*, note 17), as well as from the broader literature on emergent capabilities in large models, which documents qualitative behavioural transitions that are not predictable from performance on component tasks.

https://doi.org/10.1017/err.2026.10152

moreover, the probability distribution changes as well: agents learn and adapt (on top of strategic agents engaging in deception and multi-agent scenarios); and finally, severity depends also on agents, not on predetermined evaluations, since agents can also act on contexts (think about sending an email). Through learning and action on the environment, the agent can in principle modify the probability and severity of potential harms.

These characteristics, which are essential for traditional conceptions of risk, are disrupted essentially by sufficiently general agents. These failures, then, are not just contingent and they cannot be resolved by more data, better models, or just doing the same things better.

Thus the product-style operationalisation of risk breaks down. However, risk as such does not become inapplicable to general agents. Humans are the most general agents we know, and we apply risk concepts to them routinely. Generality then, forces a change in the level at which risk is assessed, rather than defeating risk altogether. Path 2, developed in Section IV.2, should be read in this light: it moves to the operationalisation of risk that already works for general agents.[21]

What can the regulator, on behalf of society, do? In the rest of the paper, I develop two possible paths to conceptualise agentic risk: (1) a conservative, formal framework, which attempts to quantify risk as a deviation from ideal behaviour (Path 1) and (2) a broader governance framework, based on an analogy to biological agents (Path 2).

For each, I'll discuss whether and to what extent they capture general agents and whether and to what extent they can be understood as risk-based.

Preliminarily, I can point out that the conservative path (Path 1) faces a principled ceiling: it can be made to work for narrow agents, but extending it to general agents would require solving the very problems that make those agents general. This of course does not make Path 2 necessary. One might accept the failure of traditional risk and conclude that general agents simply should not be deployed, or that some entirely different and as-yet-undeveloped framework is needed. The case for Path 2 as the appropriate response depends on a further argument I'll develop in Section IV.2.

## IV. Two paths for agentic risk

In this section, I take the definition of risk found in international standards (probability and severity of harm combined) seriously, and I study two ways to apply it to AI agents, depending on whether they are narrow (Path 1) or general agents (Path 2).

### *1. Path 1: quantifying agentic risk for "narrow" agents*

The standard definition of risk as a combination of probability and severity of harm requires two things: a way to estimate how likely harmful outcomes are and a way to

[21] This third failure connects to the older economic literature on the distinction between risk and uncertainty, mentioned in Section II above as sense (v) of "risk." On one reading of F H Knight, *Risk, Uncertainty and Profit* (Houghton Mifflin 1921), what separates quantifiable risk from genuine uncertainty is not whether the probability calculus can be applied in the abstract, but whether the underlying probabilities are *stable* enough. Probabilities attaching to natural events typically are; probabilities attaching to harms in social contexts often are not, because the behaviour of the parties responds to the assessment itself, through adverse selection and moral hazard: see S F LeRoy and L D Singell, "Knight on Risk and Uncertainty" (1987) 95 Journal of Political Economy 394, esp. p 401 on the impossibility of insuring entrepreneurship. S O Hansson, "Can Uncertainty Be Quantified?" (2022) 30 Perspectives on Science 210, makes a related point about the limits of quantification, and J Daníelsson, R Macrae and A Uthemann, "Artificial Intelligence and Systemic Risk" (2022) 140 Journal of Banking & Finance, apply it to systemic risk arising from AI in financial markets. In the case of AI agents, this also stems from the assessed system's own learning and its action on the environment, so the distribution shifts even holding the surrounding population fixed. Many thanks to an anonymous reviewer for suggesting this connection.

https://doi.org/10.1017/err.2026.10152

estimate how bad they are when they occur. For narrow agents, operating in sufficiently bounded environments, both are obtainable. The main issue is how to construct them from the materials that an agent-based system in fact provides.[22]

An agent, unlike a product, does not have a fixed input–output profile that can be tested exhaustively before deployment. What it has instead is a policy, i.e., a mapping from states to actions, and an environment in which that policy is executed. Risk, on the Path 1 approach, is therefore a property of the policy–environment pair rather than of the system in isolation: it measures how far the agent's actual behaviour deviates from what an ideally aligned agent would do in the same environment, or equivalently, how likely the agent is to violate constraints that represent the boundaries of acceptable conduct. These two formulations correspond to the two options developed below: Option 1 and Option 2, which differ in their informational requirements and regulatory tractability. The epistemic/bouletic distinction (see *supra*, note 22) maps naturally onto the two components of risk: the agent's epistemic state, combined with the environment's transition dynamics, determines the probability distribution over trajectories; the bouletic element, via the reward function or goal structure, informs which trajectories count as harmful and how severely. However, severity is assessed against an external standard – the preferences or protected interests of those affected – not against the agent's own reward signal. In fact, a misaligned agent that efficiently pursues its own objectives may register near-zero deviation from its internal ideal while producing very severe outcomes by human standards.

In the rest of this section, I sketch the two approaches. I explain them informally and leave the formal details to future work.

### *a. Option 1: preference-based*

The first option starts with a preference relation over trajectories (grounded in a formal approach to alignment developed by the author elsewhere[23]). The key idea is to define (and then measure) agentic risk as the expected gap between what an ideally aligned agent would achieve and what the actual agent does.

This is based on a number of standard (but not necessarily unproblematic) assumptions. Human evaluators hold preferences over *lotteries* of trajectories (not just deterministic outcomes) satisfying the von Neumann–Morgenstern axioms (completeness, transitivity, continuity and independence).[24] This yields a *cardinal* utility function $U$ over trajectories, unique up to positive affine transformation, against which all outcomes can be measured on a common scale. From this, one can define an optimal policy $\pi^*$ (i.e., the policy a fully aligned agent would follow) and use its utility $V^*(s_0)$ as the benchmark for any given initial state ($s_0$). The severity of a realised trajectory $\tau$ is then $V^*(s_0) - U(\tau)$, the shortfall relative to optimal expected performance. Agentic risk is the expectation of this severity over the agent's actual trajectory distribution.

[22] The philosophical tradition offers a useful framing here. To predict and evaluate an agent's behaviour, one need not have access to its internal architecture: it could be sufficient to adopt what D C Dennett, *The Intentional Stance* (MIT Press 1987), calls the "Intentional Stance," i.e., treating the system as if it has beliefs and desires, and predicting its behaviour accordingly. This stance could distinguish an epistemic component (its representation of the environment, grounding probability estimates) and a bouletic component (its goals or reward function, grounding severity assessments). The two components have opposing "directions of fit" (GEM Anscombe, *Intention* (Harvard University Press 1957); J R Searle, *Intentionality: An Essay in the Philosophy of Mind* (Cambridge University Press 1983)): beliefs aim to match the world, desires aim to change it. Risk, on this picture, arises from the interaction between what the agent believes is possible and what it is disposed to pursue.

[23] F Faroldi, *Intention and Alignment in AI Agents* (2025) (under review).

[24] One can relax these assumptions and work with a partial ordering, ending up with sets of utility functions instead, perhaps being more faithful to genuine normative uncertainty. The development of this line of thought is left to future work.

https://doi.org/10.1017/err.2026.10152

This option has some advantages: it is formally precise, it integrates naturally with the structure of Markov decision processes (MDPs), a very general mathematical framework to capture agency as sequential decision-making and it yields a single scalar risk measure that can in principle be tracked in real time. It also has costs. First, it requires strong assumptions: a complete and transitive preference ordering over trajectory lotteries, cardinality and a utility function $U$ that must be specified before deployment. Second and most importantly, it requires that the optimal policy $\pi^*$ be known in advance, which is what we argued in Section III cannot be done for sufficiently general agents.

### *b. Option 2: constraint-based*

There is a different take, which promises to overcome some of the issues just highlighted while at the same time remaining within the broad agential framework just sketched. In brief, rather than starting from a (human) preference relation that determines an ideal trajectory (and thus suffers from having to know some details in advance), we switch to a constraint-based framework. Thus, the agent's risk will be defined, and then measured, as a function on the constraints that the agent meets (or does not meet).

The first stage is *normative specification*: defining what the constraints are. Constraints of sufficient generality can be stated without knowing the specific environment in which the agent will be deployed, the particular task it will pursue, or the trajectory space that will be realised. In this sense, the constraint-based approach does not require knowing the environment in advance: the content of the normative standard is environment-independent.[25]

The second stage is *probabilistic evaluation*: calculating or estimating the probability that the agent will violate those constraints during deployment, what we may call the agential risk. This stage requires knowledge of the environment, and the constraint-based approach thus inherits the same informational demands as the preference-based one at this point.[26]

The principal advantage of Option 2 over Option 1 is the reduction of the advance specification burden of what counts as harm. But the computation of agentic risk as the expected weighted sum of constraint violations still presupposes a tractable model of how likely violations are under the agent's policy.

Let's take stock of what the formal reconstruction actually adds. If it only works for narrow agents, and narrow agents are the ones for which traditional risk assessment might have worked in the first place, is it merely an elaborate restatement of what we already had?

I think the answer is no, for at least two reasons. The first is that the formal framework offers something different from the qualitative risk-matrix approach that dominates current practice: it yields a scalar risk measure that can, in principle, be computed or estimated in real time as the agent operates. This opens the door to automated monitoring of deployed agents, something that the traditional approach, which relies on *ex ante* human judgement about intended purpose and foreseeable misuse, cannot deliver at scale. The second reason is diagnostic. The formal apparatus makes explicit what assumptions must hold for agentic risk to be well-defined at all: that the state and action spaces are tractable, that the transition dynamics are stable enough to support probability estimation and that the relevant evaluative standard (a preference ordering or a constraint set) can be

[25] One could also generate a preference relation based on such general features, thus bypassing the need for a concrete preference relation over trajectories. Under certain assumptions, the preference-based and constraint-based options are provably equivalent; see F Faroldi, *Intention and alignment in AI Agents* (2025) (under review).

[26] Estimating violation probabilities over trajectories requires either a formal model of the environment – a transition function $T$ in the MDP framework – or empirical sampling through simulation or controlled deployment.

https://doi.org/10.1017/err.2026.10152

specified in advance. The question of where those preconditions cease to hold, and Path 2 becomes necessary, is one I return to in Section V.

## 2. *Path 2: agent governance: law-following AI for "general" agents*

Path 1 was conservative: keep the notion of risk that traditionally applies to products and extend it for (a general class of) agents. Path 2, which I will sketch in this section, takes a broader perspective, i.e., investigates how to apply a notion of risk to more general agents, starting from our aims and an analogy.

Both in products and in agent management, in fact, we want to avoid harm. How do we avoid harms for, and caused by, the agents we are familiar with, i.e., biological agents? When it comes to non-human, but still biological agents, the systems we use for their management are less complex than those applied to humans and operate on two levels.

At the first, *physical* level, i.e., the direct "handling" of what the animal does, the tools are, broadly, physical means of restraint and a behaviouristic system of positive and negative reinforcement (e.g., dog training). Non-human animals have fairly well-understood mental and physical capabilities, which are not too surprising and do not change too fast for our coping mechanisms. At the second, *legal* level, we adopt a combination of strict liability, negligence-based liability and vicarious responsibility regimes for wild, dangerous animals, and domestic animals (the details obviously vary by jurisdiction but are not important for the current discussion). These mechanisms allocate the financial consequences of (non-human) animal agentic risk in a principled way. Legal liability for animal harm is addressed solely to the human principal, because the animal cannot internalise a norm or act for reasons, it cannot be a bearer of legal obligations in any meaningful sense.

The systems we use to collectively "handle" fellow human agents (e.g., to prevent, reduce or manage the harms humans can cause) are more complex. At a sufficiently abstract level, they can of course be described as performing a kind of risk-management. However, the systems in question (e.g., law, ethics, social norms, institutions) do not work by identifying an intended purpose for human beings and enumerating reasonably foreseeable misuses of them, as standard risk management would suggest. They address obligations and prohibitions (among other things) directly to the agent, on the standing assumption that the agent can understand those norms, act on them and be held responsible when it violates them. Two features (for our purposes) make this possible: we have a sufficiently reliable theory of mind for other humans to attribute intentions, beliefs and reasons to them; and we are capable of internalising normative standards rather than merely responding to incentives. These two features make normative governance the appropriate (and possibly also in the sense of justified) tool; it thus seems that it is the presence or absence of homologous features in AI agents that will determine whether the same approach can be extended to them.

But why should the features that make normative governance appropriate for humans also make it appropriate for sufficiently general AI agents, rather than the simpler alternatives of physical restraint or behaviouristic reward? Those simpler alternatives are appropriate when, and because, the agent in question lacks certain features that make them insufficient. Three such features are relevant here.

The first is *reason-responsiveness*: the capacity to modify behaviour in response to normative reasons rather than merely causal incentives, e.g., to distinguish the legitimacy of a norm from the cost of its violation, and to act accordingly.[27] This is the feature that

[27] The distinction between reasons and incentives is not merely philosophical: it has practical regulatory significance. An agent that complies with a norm only because the expected penalty exceeds the expected gain will systematically defect whenever it believes detection is unlikely. An agent that has internalised the norm as a reason will not; or at least, its failure to comply will represent a different kind of failure, one that calls for

https://doi.org/10.1017/err.2026.10152

makes normative governance possible in the first place: a system that cannot recognise reasons cannot be addressed by them, and must instead be managed through physical or incentive-based means alone.

The second is *behavioural unpredictability*: the agent's possible behaviours are too varied and context-sensitive to be exhaustively constrained by physical means or enumerated rules. This is a structural limitation: as Section III argued, the action space of sufficiently general agents is open-ended, which means that any finite set of physical or behavioural constraints will leave significant regions of that space ungoverned. Normative governance handles this through the generativity of legal reasoning: norms can be extended, e.g., by analogy to novel cases without requiring advance enumeration.

The third is *opacity of internal states*: we cannot reliably infer the agent's goals, values or degree of norm-acceptance from its observable behaviour alone. An agent that complies under observation and defects otherwise is indistinguishable, from the outside, from one that has genuinely accepted the norm; only the internal architecture can tell them apart. This is, notably, not a problem peculiar to AI: for instance, Hart insisted on the internal point of view as a necessary component of any adequate account of legal obligation.[28]

Non-human animals fail the first condition,[29] and presumably only meet the second and third in a limited way; humans meet all three, which is why their "management" requires law, ethics and other social norms rather than mere behavioural engineering. Sufficiently general AI agents (should) also meet all three: either by construction (as argued by the LFAI approach to be introduced presently) for the first condition (which is prescriptive) or by definition, as for the second condition, or by observation, for the third condition (the opacity of internal states is a well-documented feature of large-scale AI systems). The second and third conditions are descriptive. The analogy to human governance can be grounded in the same features that make normative governance the appropriate tool for managing human agents.

If the three conditions above are met, neither physical restraint nor product-style risk assessment is adequate; only normative governance is left. The law is well-suited to this role, since while it does not perform risk assessment in the product-safety sense, it operates at a higher level: it sets background rules that shape incentives and assign liability, inducing risk-minimising behaviour across a population of actors and over time, without requiring that any single party calculate the probability and severity of each harm in advance.[30] More importantly for our purposes, the law achieves this by addressing obligations directly to agents, on the assumption that they can recognise those obligations as reasons for action. Can sufficiently general AI agents be built to be the kind of thing that law can meaningfully address? That is what the Law-Following AI (LFAI) framework proposes.

different governance responses. There is growing technical work exploring how to separate reward as a causal signal from value as a normative standard; see J Svegliato, S B Nashed and S Zilberstein, "Ethically Compliant Sequential Decision Making" (Proceedings of the AAAI Conference on Artificial Intelligence, 2021), and the discussion in Section IV.2.

[28] H L A Hart, *The Concept of Law* (3rd edn, Oxford University Press 2012) pp 89–91.

[29] Though see K Andrews, "Naïve Normativity: The Social Foundation of Moral Cognition" (2020) 6 Journal of the American Philosophical Association 36, for the view that some non-human animals possess a naïve form of normativity, which would complicate the first condition at least for some species.

[30] This is perhaps the most interesting lesson of the law and economics analyses of tort law; see, e.g., W M Landes and R A Posner, *The Economic Structure of Tort Law* (Harvard University Press 1987); G Calabresi, *The Costs of Accidents: A Legal and Economic Analysis* (Yale University Press 1970); G Calabresi, "Some Thoughts on Risk Distribution and the Law of Torts" (1961) 70 Yale Law Journal 499; R A Posner, *Economic Analysis of Law* (9th edn, Wolters Kluwer 2014), which treat risk as a macro-level policy variable rather than a quantity computed ex ante about any particular actor.

https://doi.org/10.1017/err.2026.10152

The Law-Following AI framework[31] proposes a strategy for ensuring that AI agents are aligned and safe (in the AI safety sense) and ethical (in the AI ethics sense) by designing them to be law-following agents. The law, broadly speaking, is a good compromise when it comes to being the source of shared values. But how can we design AI agents that are law-following?

Autonomous agents must be built so as to adopt an internal point of view because we cannot judge their (internal) acceptance of a norm solely by observing their exterior conduct. Thus, we have to *make sure that autonomous agents act for reasons* (and for the right reasons). The internal point of view, i.e., roughly, the acceptance of a rule, provides a reason to act in a certain way that is not there when the conduct is unregulated. By reason, in this particular context, I mean not just the usual justificatory, objective reason (the standard use of "reason" in this work), but also a motivating, possessed (or subjective) reason, really, a reason for (in)action. This reason is legal, i.e., the existence and believed legitimacy of a particular legal norm.

The technical challenge that LFAI is confronted with is to integrate reasons in the architecture of (at least some) AI systems. The main problem is that the reward function encodes both technical and evaluative information, thus leading to many of the problems seen above. Namely, this function carries *instrumental* information about how the task is to be accomplished, i.e., which states represent progress towards the goal; and it carries *evaluative* information about what may permissibly be done along the way, i.e., which states and actions are off-limits and at what cost. The first step will be separating the two by adding a formal evaluative system that filters the normatively acceptable reward functions (Svegliato and others explore part of the needed solution, but do not solve the motivation question).[32] Second, reasons need to be introduced, for instance by *decomposing the reward function into factors*,[33] and having agents learn reasons (via inverse reinforcement learning).[34] The agent of course needs to be able to act for the relevant reasons, not just consider them as a cost to be avoided or weighed.[35]

While the exact implementation details are not important in this context, what matters is that (an approach like) LFAI captures a more abstract form of risk management and applies to fully general agents: agents that are at least as general as humans, since humans unquestionably engage in law-following.[36]

[31] This particular version has been developed in F Faroldi, Law-following AI (2026) under review; but ideas in the vicinity can be found in N Kolt and others, "Legal Alignment for Safe and Ethical AI" (arXiv, 2026) <https://arxiv.org/abs/2601.04175>; C O'Keefe and others, "Law-Following AI: Designing AI Agents to Obey Human Laws" (2025) 94 Fordham Law Review 57.

[32] Svegliato, Nashed and Zilberstein, *supra*, note 26.

[33] Cf., in a different context, Z Juozapaitis and others, "Explainable Reinforcement Learning Via Reward Decomposition" (IJCAI Workshop on Explainable Artificial Intelligence, 2019); F L G Faroldi, "Trasparenza dell'algoritmo e deep learning. Note logiche a margine della proposta di regolamento sull'Intelligenza Artificiale (Artificial Intelligence Act) della Commissione Europea e di un'ordinanza della Corte di Cassazione" (2021) 10 Revista Iustitia.

[34] Cf. S Russell, *Human Compatible* (Viking 2019).

[35] Let me address an objection I am grateful to an anonymous referee for pressing: does this filter not presuppose that we already know which reward functions are normatively acceptable, and does that not return us to the advance specification burden that defeated Path 1? I think not. First, in Path 1 the filter, as it were, was on preferences over trajectories; whereas here it is on objectives, which can be stated at a level of generality that does not mention the deployment environment at all. Second, here the content of the admissibility, contrary to Path 1, is supplied by an existing legal order, publicly promulgated, institutionally maintained, and equipped with its own machinery for determining what it requires in cases nobody enumerated in advance. Third, Path 1 needs its evaluative standard to be complete enough to yield a determinate scalar before deployment, whereas the LFAI filter has no such requirement: it can screen out clearly impermissible objectives while leaving hard cases to be settled downstream by the buck-stopping mechanisms discussed in Section V, which is what the legal system ordinarily does.

[36] The LFAI framework presupposes a functionalist account of agency and norm-following: what makes an agentic system capable of following norms is constituted by its functional organisation, not by its physical substrate or intrinsic properties. An objection could be that genuine law-(and more generally, rule-) following

https://doi.org/10.1017/err.2026.10152

Before comparing the two paths, let me address three objections to Path 2.

The first concerns the status of reason-responsiveness. Of the three conditions above, the second and third are descriptive, whereas the first is prescriptive: general agents should be built to be reason-responsive. This gives the argument a conditional structure. The claim that normative governance is the appropriate tool presupposes agents that satisfy the very condition that makes it appropriate, and whether such agents can be built is an open technical question, the one the LFAI research programme addresses. The argument of this section is therefore conditional: if agents can be built to satisfy all three conditions, normative governance is the appropriate framework for them; if they cannot, then neither path applies to general agents, and the case for restricting their deployment altogether becomes correspondingly stronger.

The second objection concerns what addressing obligations to the agent accomplishes, if to enforce obligations sanctions are needed, and sanctions will fall on the human principals, if any. A sceptic will conclude that Path 2 collapses into the animal model with better training: norms addressed, in the end, only to the humans who pay.[37] The main reply is that internalisation does governance work that liability allocation alone cannot do. Liability operates *ex post* and through the principal's incentives, since it disciplines the choice of which agents to deploy, but it cannot guide the deployed agent's conduct in the novel, unmonitored situations that Section III identified as the distinctive locus of agentic risk. A law-following agent can extend legal standards to cases no one enumerated, in the way legal reasoning is designed to be extended. In the animal model, the generativity of law is available only to the humans around the animal; in LFAI, it is available to the agent itself.

The third objection is, in my mind, the most serious. Design-time instillation of the internal point of view looks, from the observer's point of view, like behaviouristic shaping just one step removed: what distinguishes an agent trained to treat norms as reasons from an agent trained to behave as if it did? This distinction may be unverifiable by behavioural means, since the two agents coincide on any finite set of observations. The problem, as already noted, is familiar from the human case: we cannot verify the internal point of view in humans either, and legal systems combine a defeasible presumption of norm-acceptance with monitoring, and sanctions for the cases where the presumption fails. A parallel architecture for AI agents, presumptive trust in certified law-following designs combined with monitoring calibrated to the stakes, seems a reasonably natural transposition.

## V. Comparison and discussion

The two paths developed in the preceding section share a common motivation: both attempt to extend meaningful risk-theoretic concepts to agentic systems that resist characterisation as ordinary products. Path 1 is more conservative: it starts from the definition of risk we find in relevant legislation (e.g., the EU AI Act), and it focuses on "narrow" agents, which can be conceptualised as having specific aims and living in

requires something more than functional organisation: phenomenal consciousness, intentionality or some other non-functional property that current AI systems lack. If this were the case, the AI would only simulate law-following, and the internal point of view could not be reached by design, as it were. I am happy to make Path 2 conditional on this functionalist assumption.

[37] The question of legal subjectivity for artificial agents has its own history, naturally: the European Parliament contemplated a form of "electronic personhood" in its 2017 resolution on civil law rules on robotics (European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)) [2018] OJ C252/239), and V A J Kurki, *A Theory of Legal Personhood* (Oxford University Press 2019), develops a general theory of legal personhood that treats animals and artificial entities within a single framework.

https://doi.org/10.1017/err.2026.10152

**Table 1.** Five dimensions of comparison between Path 1 (quantified agentic risk) and Path 2 (agent governance). The two paths are not competing answers to the same question but complementary frameworks operating at different levels of abstraction.

| Dimension | Path 1: quantified agentic risk | Path 2: agent governance |
|---|---|---|
| **Precision vs. generality** | High formal precision; restricted to narrow agents with tractable policy spaces | High generality; applies to agents as general as humans, at the cost of quantitative tractability |
| **Completeness** | Requires harm space and preferences/ constraints specified in advance; reintroduces "intended purpose" formally | Defers incompleteness to legal interpretation; handles novel cases through analogy and evolving doctrine |
| **Point of view** | Third-person: risk computed externally by measuring trajectory distributions against a fixed standard | First-person: agent must adopt the Hartian internal point of view, accepting norms as genuine reasons for action |
| **Regulatory burden** | Falls on the provider/deployer to specify, certify and update constraint sets or preference relations | Distributed between designer (law-following architecture) and agent (norm internalisation); requires new institutional mechanisms |
| **Harm taxonomy** | Requires a closed-world enumeration of harm categories {Ck}; formal elegance rests on a bounded harm space | Open-ended; legal reasoning generates judgements on novel harms without prior enumeration |

controllable environments. Path 2 is more ambitious: it proposes to consider agents as general as humans, and it adopts a much more abstract understanding of risk. In this section, I evaluate the approaches sketched above along two axes: whether and to what extent they capture agency and whether and to what extent they are risk-based. I organise the comparison along five dimensions, which are summarised in Table 1.

### *1. Generality*

The most important structural difference between the two paths is that each achieves its primary virtue by sacrificing the primary virtue of the other: Path 1 attains precision by restricting the class of agents to those that can be formally modelled with stable preferences; Path 2, by contrast, gains generality by abandoning per-system quantification of risk in favour of normative governance. It is therefore less precise, but more widely applicable. The choice between the two paths depends on whether the agent in question is "narrow enough" for Path 1's formal preconditions to hold. Throughout this paper, I have treated the narrowness/generality distinction as given, but the usefulness of the two-path architecture depends on being able to tell, for a given agent in a given deployment, whether it is general (or narrow) "enough."

One connection to existing work is important here. The dimensions of agency surveyed in Section II – autonomy, generality, goal complexity and the like – correlate with the likelihood that these formal preconditions will fail. An agent with high autonomy and broad generality is very likely one whose action space resists enumeration and whose transition dynamics are environment-dependent. The multidimensional accounts of agency proposed by Kasirzadeh and Gabriel and by Dung could in principle be connected to these formal preconditions.[38]

[38] Kasirzadeh and Gabriel, *supra*, note 7; Dung, *supra*, note 7.

https://doi.org/10.1017/err.2026.10152

### 2. Completeness

Path 1 reintroduces a version of intended purpose, via the notion of an ideal trajectory (an optimal policy against which deviation is measured) or a set of constraints, both of which need to be known in advance. If the critique of Section III – that intended purpose cannot be specified for sufficiently general agents – is right, it then applies equally to the formal framework of Path 1. Moreover, probability estimation requires a sufficiently closed and modellable environment, again contrary to what was established in Section III.

Path 2 tolerates a degree of open-endedness that suits general agents very well: not everything can and has to be specified in advance. This comes at (at least) two costs: a degree of imprecision or uncertainty, and the necessity of buck-stopping mechanisms (such as traditionally, supreme courts).

### 3. Point of view

Path 1 treats risk as a third-person, externally observable quantity. Path 2, by contrast, requires the agent to adopt what Hart called the internal point of view: the genuine acceptance of a norm as a reason for action, rather than merely a constraint on behaviour whose violation triggers a sanction. Path 1 could be implemented as an external audit mechanism, which seems pretty similar to what is happening with benchmarks and evaluation efforts. Path 2 instead requires a novel architecture built into the agent's practical reasoning, which seems a lot harder to ascertain (exactly like in the human case).

### 4. Regulatory burden

Path 1 imposes the regulatory burden on the provider or developer, in line with the EU AI Act, which requires providers of high-risk systems to maintain technical documentation and conduct systematic risk evaluations prior to market placement, including the system design in some cases. This path is more directly applicable in the context of the existing regulatory framework. Path 2 surely requires that the developer implement the appropriate law-following architecture, but it also includes the agent as some sort of legal subject. This implies new institutional mechanisms without clear analogues in current product-safety law: something like legal standing for AI agents or liability regimes calibrated to degrees of autonomy. Any workable implementation will require mechanisms that ensure the financial consequences of agentic harm fall on parties who can bear them, such as mandatory insurance, developer backstop liability, sector-specific restitution funds, or a combination thereof.

### 5. Harms

Analogously to the second consideration, for Path 1, harms seem to have to be individuated in advance, either concretely or abstractly; for Path 2, there is no such requirement: the usual mechanism of the law (analogy, extension, etc.) can be applied to, e.g., new harms.

## VI. Conclusion and future work

Traditional risk assessment frameworks assume a fixed intended purpose and reasonably foreseeable misuse. General-purpose AI agents, by contrast, do not fit these assumptions. They may lack a clearly defined purpose and are capable of unanticipated, creative or emergent behaviours that challenge the notion of "foreseeability." The current legislation conceptualises AI systems as products, aligning risk with intended purpose.

https://doi.org/10.1017/err.2026.10152

This framework may be insufficient or inadequate for dealing with agents who operate with a degree of autonomy and unpredictability. While systems exist to manage risks posed by biological agents (e.g., laws, institutions, reward/punishment mechanisms), these are not fully translatable to artificial agents – particularly those whose internal states or goals we cannot directly access or influence. Especially relevant is the scenario in which an agent is not currently aligned but could be, under counterfactual conditions. This introduces the possibility of developing quantitative definitions of risk grounded in optimal trajectories and expected behaviour deviations. However, in the two cases we considered, the narrowness of agents allowed us to get a good grip on being precise about agentic risk; whereas generality forced us to abandon the per-system, *ex ante* operationalisation of the standard notion of risk in favour of the population-level operationalisation that the law already embodies.

This trade-off suggests a policy implication: risk-based regulation, such as the current EU AI Act, may be adequate for narrow agentic risk (along these conceptual lines, while it could still be inadequate when it comes to other considerations), whereas a broader regulatory framework or an ampler definition of risk, is needed for truly general agents. The comparison of Section V identifies preliminary conditions under which Path 1's formal apparatus applies and suggests that the dimensions of agency familiar from the recent literature (e.g., autonomy, generality, goal complexity) correlate with the likelihood that those conditions will fail. The boundary between narrowness and generality remains the central open problem. Like the question of what constitutes intelligence, the boundary keeps shifting as our modelling capacities improve and as agents themselves become more capable. Developing principled criteria for locating this boundary in concrete regulatory contexts, and understanding how the two paths can be integrated within a single governance architecture, are the most pressing tasks ahead.

**Acknowledgements.** Versions of this work have been presented or discussed in Pavia, Bochum (online workshop), Berkeley and Cambridge. I am grateful to the many people who engaged with the present work, chiefly Luca Zanetti.

Claude (Anthropic) Opus 4.6, 4.8 and 5 were used during 2025 and 2026 for clarity and grammar, for generating part of the LaTeX code, transforming the TeX to docx and for formatting the bibliography. The tool is publicly available at https://claude.ai. It was used interactively, on drafts written by the author; no output was incorporated without the author's review and revision, and all substantive claims, arguments and references are the author's own.

**Competing interests.** The author declares none.

**Cite this article:** FLG Faroldi, "Risk for AI Agents". *European Journal of Risk Regulation*. https://doi.org/10.1017/err.2026.10152

https://doi.org/10.1017/err.2026.10152