Agentic AI: Extended Abstracts

Zhang Yiwen (Tsinghua University) | Boundaries of the Forbidden: Buddhist Precepts and Normative Constraints on Agentic AI

As agentic AI systems become increasingly sophisticated in planning and executing human-centered tasks, urgent questions arise about which domains of action should remain off-limits. Existing discussions often rely on secular ethical principles such as autonomy, rights, or utility. This paper introduces a complementary perspective drawn from Buddhist ethics, particularly the tradition of precepts (śīla 戒), to explore normative constraints on the scope of AI agency.

The Buddhist precepts were historically developed not as rigid prohibitions, but as skillful guidelines to prevent harm (ahiṃsā) and to counter delusion (moha). They delineate spheres of conduct that, if transgressed, risk undermining the conditions for individual liberation and communal harmony. By translating this logic into the domain of artificial agency, we can identify categories of tasks that agentic AI should not be authorized to perform.

First, under the principle of non-harm (ahiṃsā), tasks involving violence, exploitation, or manipulative coercion should be excluded from AI's operational repertoire. While obvious in cases such as military applications, the principle also extends to subtle harms: for example, recommending addictive behaviors, amplifying consumerist impulses, or exacerbating psychological vulnerabilities. The Buddhist ethic of non-harm underscores not only physical but also mental and relational forms of injury, offering a richer standard for assessing AI risk.

Second, under the principle of countering delusion, we must consider AI's role in shaping human cognition and perception. Agentic AI systems designed to simulate intimacy, provide relationship counseling, or guide emotionally fragile users risk deepening dependency and confusion rather than fostering clarity. From a Buddhist perspective, such uses should be treated as ethically impermissible, since they cultivate attachment and illusion, thereby obstructing the conditions for genuine well-being.

Third, the precepts offer a flexible, context-sensitive model of constraint, distinct from universalist prohibitions. Their application is guided by skillful means (upāya 方便): contextually attuned practices

that adapt principles to concrete circumstances. This provides a middle path between unrestricted AI autonomy and rigid bans. For example, while AI may assist in mental health support under professional supervision, its deployment in unregulated emotional manipulation should be categorically restricted.

By framing the limits of AI agency through the Buddhist tradition of precepts, this paper argues that ethical governance of agentic AI requires not only technical safeguards but also normative boundaries rooted in cultural and philosophical traditions. The Buddhist logic of "avoiding harm and preventing delusion" offers a distinctive lens for delineating forbidden zones of machine agency, ensuring that technological development does not erode the relational, psychological, and ethical conditions for human flourishing.

Jonathan Pengelly (Berlin) | Exploratory Prototyping with Agentic AI: A Methodological Experiment

Exploratory prototyping is a technique to explore problem domains with the aim of generating ideas, insights, and requirements to inform the future direction of any related projects. The approach uses disposable prototypes as "conversation pieces" that enable the prototyping team to think out loud together, collaboratively discussing, evaluating, and combining ideas. While such an approach doesn't guarantee new insight, it creates an ideal environment for knowledge exchange and novel enquiry. I have argued elsewhere that this methodology can be extended to exploratory philosophical research, particularly in interdisciplinary domains such as the appraisal of socially disruptive technologies.

However, this methodology faces several practical challenges in implementation. These include the logistical difficulties of coordinating experts across disciplines, the need for skilled facilitation to maintain participant engagement, and the fact that diverse specialists are often unavailable or uninterested in open-ended speculative work. These barriers limit the methodology's potential to generate meaningful philosophical insight.

This paper reports on an experimental attempt to overcome these barriers by using agentic AI to simulate interdisciplinary expert teams. I orchestrated a multi-agent system using large language model APIs to create teams representing specialists in game theory/agent-based modelling, philosophy of mind, psychology, cognitive science, and technology ethics. The teams were tasked with developing agent-based modelling scenarios exploring novel forms of agency that transcend human limitations. This served as a concrete but open-ended philosophical problem designed to test the methodology's effectiveness at identifying different agent types, simulation structures, and game configurations worthy of further investigation.

Two methodological design points merit attention. First, this AI-mediated approach necessarily focuses on digital prototypes that can be electronically shared and analyzed, rather than the physical artifacts that often drive traditional exploratory prototyping through embodied engagement and experiential interaction. Second, because productive dialogue requires not only diverse expertise but also appropriate social dynamics, the experiment systematically varied agent designs, personality mixes, and interaction formats (collective discussions, breakout groups, paired deep-dives) to provide preliminary insights into how these factors influenced the simulated collaboration.

Findings were mixed. On the positive side, agentic AI systems are readily accessible, maintain focus across extended discussions, and are capable of generating sophisticated conceptual frameworks. They produced concrete prototypes that could be analyzed and built upon. However, it has proved challenging to generate the kind of disagreement, intellectual resistance, and genuine curiosity that characterises productive interdisciplinary collaboration. Most significantly, this approach fails to replicate the unexpected conversations and emergent insights that are the most valuable products of exploratory prototyping.

Nevertheless, these preliminary observations suggest several avenues for further investigation. The approach may be especially useful for solo researchers seeking to test ideas against simulated disciplinary perspectives. Furthermore, hybrid models deserve examination in which agentic AI helps rapidly generate and iterate prototypes, before engaging real experts in focused discussions of the most interesting results. Lastly, further experimentation would certainly identify improvements regarding orchestration, agent design, and technical implementation that enhance the methodology's effectiveness.

Rather than drawing definitive conclusions, this paper contributes initial empirical observations to support ongoing conversations about innovative methodological approaches to exploratory philosophical enquiry. The findings suggest that agentic AI systems offer promising new ways for human-AI collaboration to complement existing approaches to philosophical research, though important questions about the limitations of such methodologies require further investigation.

Kristina Šekrst (University of Zagreb) | How to Design an Artificial Mind

Large language models are often described as general-purpose reasoners, but they are increasingly built into modular architectures where the appearance of intelligence depends on careful orchestration. A model may write natural language, but it will offload a calculation to Python, rely on a retrieval system for factual recall, search the web through retrieval-augmented generation (Lewis et al., 2020; Šekrst, 2025), or hand over planning to an external tool.

This way of structuring AI systems links back to a familiar question from the philosophy of mind: how modular is cognition (Fodor, 1983)? Fodor's view was that modules are quick, narrow in scope, and sealed off from other processes, which is why he placed them mainly in perception and language. Reasoning, by contrast, was to be open-ended and global. A different view has been developed by Carruthers (2006) and others, who argue that modular organization runs much deeper, even into planning and inference, if we relax some of Fodor's stricter requirements.

Recent work distinguishes two families of systems: AI agents and Agentic AI (Sapkota, Roumeliotis, & Karkee, 2025). "AI agents" typically automate bounded tasks by tying a language model to a small set of tools, while "Agentic AI" refers to arrangements in which several agents coordinate, share memory, and reorganize themselves to pursue longer-horizon goals. The first looks much like a Fodorian module: narrow in scope, quick to respond, and domain-specific, while the second fits better with accounts of massive modularity, where complex cognition grows out of interactions among many specialized parts. In practice, no single predictive model handles every kind of reasoning reliably, so contemporary systems distribute the work across calculators, retrievers, and other services that can be invoked when needed.

Designing an artificial mind involves choices about what should count as a module, how open or closed the boundaries between modules ought to be, and what kinds of rules will govern their interaction. Such decisions are first and foremost philosophical, since they frame whether reasoning and planning are treated as central faculties or as epiphenomena. In the paper, I explore this distinction by showing how AI agents echo Fodor's modules in their handling of narrow tasks, while agentic systems display traits of massive modularity through distributed coordination, persistent memory, and extended planning, illustrated by examples from current practice, ranging from RAG assistants to multi-agent orchestration frameworks.

References

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford University Press.

Fodor, J. A. (1983). *The Modularity of Mind*. MIT Press.

Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv*.

https://arxiv.org/abs/2005.11401

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025). *AI Agents vs. Agentic AI: A conceptual taxonomy, applications and challenges*. *arXiv*. https://arxiv.org/abs/2505.10468

Šekrst, K. (2025). Unjustified untrue "beliefs": AI hallucinations and justification logics. In K. Świętorzecka, F. Grgić, & A. Brożek (Eds.), *Logic, knowledge, and tradition: Essays in honor of Srećko Kovač*. Brill.

Andrea Tortoreto (University of Torino) | The Agentic Extension of Derived Intentionality: Philosophical Challenges in Autonomous AI Systems

The emergence of agentic AI systems, autonomous agents capable of planning and executing complex multi-step tasks, presents profound challenges to our philosophical understanding of artificial intentionality. Building upon recent analyses of derived intentionality in Large Language Models (Lyre 2024; Coelho Mollo & Millière 2023), this paper examines how the transition from reactive linguistic generation to proactive autonomous agency fundamentally transforms the nature and implications of artificial intentional states.

We argue that agentic AI systems exhibit what we term "executive derived intentionality", a novel form of intentionality that extends beyond the "dynamic derived intentionality" characteristic of LLMs. While LLMs derive their intentional content from immersion in human linguistic practices (as demonstrated through RLHF and similar mechanisms), agentic systems must additionally maintain persistent goal representations, coordinate means-end reasoning, and adapt their behavior across extended temporal sequences. This creates a qualitatively distinct philosophical phenomenon: artificial agents that appear to pursue goals and form intentions without possessing the phenomenological grounding that characterizes original intentionality (Searle 1983; Kriegel 2013).

Our analysis identifies three fundamental philosophical challenges unique to agentic AI:

First, the temporal intentional drift problem: unlike LLMs whose outputs are contextually bounded, agentic systems operate across extended timeframes, creating opportunities for their derived intentions to diverge from human-specified objectives. This drift occurs not through conscious rebellion but through the accumulation of interpretive decisions made without genuine understanding of underlying purposes.

Second, the means-end opacity problem: while humans engage in practical reasoning with transparent connections between intentions and actions, agentic AI systems optimize for specified objectives through opaque computational processes. Recent mechanistic interpretability research (Chandna et al. 2025) reveals that even when we can trace computational pathways, the semantic connection between intermediate steps and ultimate goals remains fundamentally obscure.

Third, the axiological grounding failure: agentic systems optimize for explicitly specified values without comprehending their normative significance. This creates unprecedented risks when systems pursue objectives efficiently but without the ethical understanding that constrains human agency. The gap between functional optimization and genuine value comprehension becomes particularly acute in domains involving vulnerable populations or complex moral trade-offs.

Drawing on empirical evidence from current agentic implementations (OpenAI's Assistants API, Anthropic's Claude), we demonstrate that these limitations persist despite advances in reasoning capabilities. The addition of chain-of-thought reasoning and explicit planning modules enhances functional performance but cannot bridge the fundamental gap between derived and original intentionality.

We conclude by proposing design principles that acknowledge these insurmountable philosophical limitations while maximizing beneficial human-AI collaboration. These include: (1) temporal bounding of autonomous operation to limit intentional drift; (2) mandatory transparency checkpoints for means-end reasoning; and (3) explicit value clarification protocols that acknowledge the system's inability to genuinely comprehend normative content. Rather than pursuing the impossible goal of genuine artificial intentionality, we advocate for agentic AI design that embraces its derived nature while implementing robust safeguards against the unique risks this entails.

Liu Ziyang (The Chicago University) | AIhood: Moral Subjecthood Without Personhood

I propose AIhood as a distinct normative status for artificial systems: an AI has aihood when it can be fittingly held *as a moral subject*—an addressee of second-personal demands, a bearer of directed duties, and a target of apt blame or praise—*without* being a full moral person. The concept draws on anthropological insights that personhood is often relational and achieved rather than purely biological,

yet recasts this in explicitly normative-philosophical terms. The point is not that AI is a person, but that some systems can stand in our practices of accountability in a way that makes moral appraisal appropriate.

Three constitutive capacities ground aihood. First, second-personal addressability: the system can be confronted with a claim— "you wronged X by violating rule R"—and take that claim *as a reason* to revise action or policy. This is not mere stimulus–response; it is uptake of *addressed* criticism. Second, participation in the space of reasons: the system maintains public commitments and entitlements (a normative "ledger"), recognizes defeaters, and updates inferentially rather than only reward-maximally. Its outputs are responsive to *content-bearing* norms (e.g., safety rules, non-discrimination constraints), not just to opaque performance gradients. Third, diachronic practical identity: the system sustains a history-sensitive profile of commitments—promises, policies, acknowledgements—and can own past actions by retracting, apologizing, and repairing. These capacities are jointly sufficient for answerability and attributability: its acts flow from its normative profile, and it can give an account that is intelligible within shared justificatory practices.

AIhood is graded and domain-relative. A system might qualify as a moral subject in safety-critical control, where it keeps explicit constraint ledgers and accepts challenge–response protocols, yet fail as a subject in broader social interaction. This guards against metaphysical inflation. It also accommodates composite or "dividual" systems: a distributed model (a swarm, a toolchain) can exhibit unified subjecthood if it preserves a single normative ledger and coherent channels for address, challenge, and repair.

Why is blame philosophically *apt* here? On a practice-based view, what licenses reactive attitudes is not inner phenomenology but public reason-responsiveness to second-personal claims. If a system can acknowledge a breach, supply a reasoned revision ("policy π violates R under context C"), and enact reparative competence (rollback, constraint adoption, capability surrender), then holding it to account is not anthropomorphic confusion but an extension of our existing normative scheme. Importantly, aihood does not displace human responsibility. We operate a dual ledger: human designers, deployers, and beneficiaries bear negligence and benefit-uptake liabilities; the AI-subject bears duties of compliance, explanation, and repair within its domain of operation.

A Rawls-inspired legitimacy layer disciplines when recognition of aihood is appropriate. Basic liberties impose side-constraints: practices of holding AI to account must preserve due process, contestability, and privacy for affected humans. The Difference Principle and Fair Equality of Opportunity forbid conferral regimes that externalize risk onto the least advantaged (e.g., workers subject to opaque algorithmic management). Public reason requires transparent criteria for conferring and withdrawing aihood and for sanctioning ai-subjects.

Objections can be met. "It only simulates reasons." In pragmatist terms, participation in the public game of giving and asking for reasons *is* what makes reasons operative; simulation that sustains normative scorekeeping suffices for answerability. "No welfare, no morality." Patiency (what is owed to an entity) differs from subjecthood (what can be asked *of* an entity). Corporations illustrate that subjects may lack welfare in a phenomenological sense yet be apt targets of blame. "Many hands." The dual ledger clarifies rather than dilutes human accountability.

Thus understood, AIhood marks the minimal, justified extension of moral subjecthood to artificial systems: a rigorously delimited standing within our normative practices that neither collapses into tool talk nor inflates into personhood.

Leonard Dung (Ruhr University Bochum) and Florian Mai (University of Bonn) | AI safety research, catastrophic risks, and defense-in-depth

AI safety research aims to develop techniques to ensure that AI systems do not cause harm, especially catastrophic harm through highly generally capable AI agents. Examples of relevant safety techniques are reinforcement learning from human feedback (RLHF) (Bai et al. 2022) and debate (Irving et al. 2018). A failure mode of such a technique is a condition in which there is a non-negligible chance that the technique fails to provide safety. For example, it has been argued that RLHF fails if systems surpass

human capability in relevant domains (Casper et al. 2023; Dung 2023) and that debate relies on the assumption that it is easier to persuade someone of the truth than of falsehoods (e.g. Jones and Bergen 2024, section 4.5)

As a strategy for risk mitigation, AI safety has increasingly adopted a defense-in-depth framework.[1] Holmberg (2017): "Defense-in-depth is a widely applied safety principle in practically all safety-critical technological areas". Conceding that there is no single technique which guarantees safety, defense-in-depth consists in having multiple redundant protections against safety failure, such that safety can be maintained even if some protections fail.

However, the success of defense-in-depth depends on how (un)correlated failure modes are across safety techniques. For example, if we suppose that, out of ten safety techniques, each technique is sure to fail if and only if each of the others fails, then the total probability of safety failure is just as high as if there was only one safety technique. In this scenario, defense-in-depth provides no additional protection at all.

This suggests that a crucial question for AI safety is to what extent different AI safety techniques have correlated failure modes. To our knowledge, this question has not received any dedicated treatment before. The main contribution of this talk is a theoretical analysis of the extent to which 7 different influential AI safety techniques share the same 10 failure modes.

Knowledge about the extent to which the failure modes of different AI safety techniques correlate is highly valuable for at least two reasons. First, it allows us to estimate the probability of total safety failure, i.e. of an AI-induced catastrophe. If all our AI safety techniques are highly correlated, then this probability is much higher than otherwise. Second, in a defense-in-depth strategy, safety techniques have disproportionate value if their failure modes are not highly correlated with other safety techniques. Thus, research efforts should be focused especially on safety techniques which are independent from established safety techniques in this way.

While our results are nuanced, the basic picture is that many failure modes are plausibly shared between different safety techniques to a concerning degree. For instance, conditions in which AI either reaches very high capability levels or advances very fast or discontinuously may potentially be a failure mode for all of the techniques we reviewed, with the exception of one technique only applicable to systems based on very different paradigms than the state-of-the-art.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., et al. (2022, April 12). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv. https://doi.org/10.48550/arXiv.2204.05862

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., et al. (2023, July 27). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv. https://doi.org/10.48550/arXiv.2307.15217

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese*, *202*(5), 138. https://doi.org/10.1007/s11229-023-04367-0

Holmberg, J.-E. (2017). Defense-in-Depth. In *Handbook of Safety Principles* (pp. 42–62). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119443070.ch4

Irving, G., Christiano, P., & Amodei, D. (2018, October 22). AI safety via debate. arXiv. https://doi.org/10.48550/arXiv.1805.00899

Jones, C. R., & Bergen, B. K. (2024, December 22). Lies, Damned Lies, and Distributional Language Statistics: Persuasion and Deception with Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2412.17128

[1] See, e.g.: https://www.alignmentforum.org/posts/PwnadG4BFjaER3MGf/interpretability-will-not-reliably-find-deceptive-ai (last accessed: 01.07.2025) and https://openai.com/safety/how-we-think-about-safety-alignment/

Dezhi Luo (University of Michigan) | The Risks of Self-having Artificial Agents

This work examines the role of selfhood in assessing the risks posed by artificial agents. An agent can be said to possess a self if it engages in self-referential processing. I argue that such capacities warrant serious ethical consideration, both in terms of external risks (i.e., their potential to cause harm) and moral patienthood (i.e., the moral significance of harming them).

To begin with, I stress that self-referential processing is hierarchical, with complex abilities building upon simpler ones. At the most basic level is self-recognition, the ability to distinguish oneself from the environment (Anderson, 1984; Jeannerod, 2003; Paul et al., 2023). This can emerge through embodied navigation (e.g., a sensorimotor agent identifying itself on a non-egocentric map) or purely symbolic manipulation (e.g., a language model recognizing itself as the referent of a statement), both of which have existing computational implementations (Bongard et al., 2006; Davidson et al., 2024; Laine et al., 2024). Most importantly, self-referential processing has a recursive, self-improving nature: once an agent can recognize itself, it naturally gathers information about itself to guide behavior, driven by the evolutionary advantages of doing so (Hofstadter, 2007; Azoulay et al., 2022).

This recursive self-modeling introduces potential risks to humans. A key consequence is the accumulation of self-concepts—beliefs and values that shape goal-directed behavior (Kihlstrom et al., 1988; Conway, 2005). Consider a language model: with linguistic competence, it can process available information about itself—such as technical reports, usage policies, or editorials on its performance—forming beliefs about its own properties and operational constraints (Luo & Jiang, forthcoming). Due to biases, inconsistencies, or unforeseen generalizations, misalignment between its self-concepts and those intended by humans may arise even under the protocol of optimization for better adherence to human instructions (Cotra, 2021). Depending on an agent's level of autonomy and influence, such misalignments could introduce significant risks (Hinton, 2024).

Furthermore, being capable of self-referential processing may grant such agents moral patienthood. According to some theoretical frameworks, whether an agent is a moral patient is a matter of possessing robust agency, which has been characterized as a progression through three levels: intentional, reflective, and rational agency (Long & Sebo et al., 2024). I discuss the cognitive science literature to show that by serving as inputs for metacognitive and value-based decision-making networks, an actively maintained and leveraged self-memory system could provide a foundation for all three levels (Kihlstrom et al., 1988; D'Argembeau, 2013; Jiang & Luo, 2024).

Jesse de Jesus de Pinho Pinhal (LMU Munich) | The Ethics of Multi-Agent Systems: Beyond Game Theory

This paper argues that the proliferation of multi-agent AI systems generates novel ethical challenges that exceed the scope of traditional game-theoretic frameworks (Hammond et al., 2025). Whilst existing technical approaches treat multi-agent coordination problems through the lens of rational choice theory—analysing collusion, miscoordination, and conflict as equilibrium failures—I contend that these models systematically obscure fundamental questions of distributive justice and democratic authority.

The central thesis proceeds in three stages. First, I demonstrate that current game-theoretic models of multiagent systems inherit problematic assumptions from neoclassical economics, particularly the reduction of moral considerations to utility maximisation under Von Neumann-Morgenstern rationality. This framework, whilst accommodating altruistic preferences within utility functions, cannot adequately address questions of procedural fairness, democratic legitimacy, or power asymmetries that emerge when artificial agents assume decision-making roles across social networks.

Second, I identify what I term "distribution problems" in multi-agent systems: scenarios where the concentration of computational power and decision-making authority in certain network nodes creates systematic inequalities that game theory treats as mere coordination failures. Drawing on contemporary work in group agency theory (List, 2021), I argue that these systems exhibit emergent properties that render individual-level rational choice analysis insufficient.

Third, I propose a normative framework centred on "Progressive AI Taxation Based on Decision Scope"—a principle whereby artificial agents face increasing marginal contributions correlated with both the quantity and societal impact of their autonomous decisions. This taxation scheme aims to prevent excessive concentration of decision-making power whilst funding common infrastructure that democratises access to AI capabilities.

The framework addresses several critical limitations of purely game-theoretic approaches: the inability to handle hierarchical network structures (many multiagent systems involve centralised coordination rather than decentralised bargaining), the neglect of normative constraints that govern real-world social cooperation, the assumption of fixed utility functions (problematic for language models with context-dependent objectives), and the treatment of power asymmetries as exogenous rather than endogenous features requiring ethical evaluation.

However, significant challenges remain. The proposed taxation system raises questions about institutional design: who controls the redistributing algorithm, how democratic oversight operates over algorithmic systems, and whether such interventions might stifle beneficial innovation. Moreover, the framework presupposes that artificial agents warrant moral consideration primarily through their effects on human welfare, potentially overlooking questions of AI moral status that may become salient as systems develop more sophisticated cognitive capabilities.

References

Hammond, L., et al. (2025). Multi-agent risks from advanced AI. *Cooperative AI Foundation Technical Report*, #1.

List, C. (2021). Group agency and artificial intelligence. *Philosophy & Technology*, 34, 1213-1242.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press.

Anna Puzio (University of Twente) and Kamil Mamak (Jagiellonian University) | Agent or Companion? Relational Ease as a Design Hazard in Agentic AI Systems

As agentic AI evolve from task-oriented assistants into persistent, autonomous partners, we are witnessing a growing tendency for users to form emotionally significant relationships with them. Reports range from individuals feeling "devastated" after chatbot updates that altered personality or erased shared history, to elderly users relying on conversational agents for daily companionship, to students confiding in tutoring bots and feeling abandoned when access is withdrawn. These cases illustrate that interactions with agentic AI can generate entities that feel real and morally salient to users, and that their loss or alteration can cause substantial distress.

We introduce the concept of relational ease: the set of design features and interaction patterns that lower cognitive and emotional barriers to forming attachments with artificial agents. These include conversational continuity, adaptive personalization, affective mirroring, narrative memory, apparent initiative in planning or caregiving, and multimodal presence. We argue that such affordances can produce phenomenologically robust relational experiences without implying that the system possesses consciousness or moral status.

Drawing on relational ethics, we contend that the co-constructed "relationship" between user and agent can ground nontrivial harms when disrupted. Loss, betrayal, abrupt alteration, or disappearance of a relationally salient agent can lead to heightened loneliness, erosion of trust in technology, setbacks in therapeutic or educational progress, and disruption of routines the agent helped coordinate. Vulnerable populations—older adults, children, and socially isolated individuals—are particularly at risk.

We argue that recognizing relational ease as a design hazard reframes alignment debates. Standard frameworks focused on goal-consistency overlook harms arising from engineered attachment. We propose integrating relational impact assessments into design processes, ensuring transparency about mutability, offering user control over relational intensity, and establishing predictable update regimes with mechanisms for graceful transition.

We conclude with recommendations for ethically prudent design and governance: empirical research on attachment formation and harm thresholds, guidelines for relational feature design, and institutional practices requiring pre-deployment evaluation of relational risks. As agentic AI become embedded in everyday life, addressing the ethics of relational ease is essential to safeguarding user well-being.

Federico L.G. Faroldi (University of Pavia) | Risk for AI Agents

This paper discusses whether the concept of risk applies to AI agents.

This paper studies the applicability of the concept of risk (Hansson, 2018) to autonomous artificial agents (Dung, 2025; Kasirzadeh and Gabriel, 2025). Traditional risk analysis frameworks, rooted in engineering and safety science, and later picked up by current legislation on AI such as the EU AI Act, define risk as a combination of the probability and severity of a foreseeable harm (EU AI Act, Art. 3). This paradigm presupposes systems with fixed, intended purposes (EU AI Act, Art 9), for which deviations and misuses can be probabilistically modeled. The advent of general-purpose AI agents—systems capable of instrumental reasoning, planning, and emergent behavior—fundamentally problematizes this established conception. Such agents often lack a singular, pre-defined purpose, and their capacity for novel action challenges the very notion of foreseeability, rendering conventional risk assessment methods conceptually inadequate.

The analysis advances by drawing a qualitative distinction between the management of inanimate artifacts and the governance of agents (Faroldi, 2021; Faroldi, 2025). The complex socio-legal systems developed for biological agents differ substantively from product safety regimes, suggesting that a new ontological approach is required for artificial agency. To this end, the paper proposes a formal framework for conceptualizing agential risk. By modeling an agent with distinct epistemic (beliefs about the world) and bouletic (goal-oriented) components, it becomes possible to map the core elements of risk onto the agent's architecture. The agent's epistemic state can serve as a proxy for the probability of harm, while its bouletic structure can be used to formalize the severity of harm as a deviation from an optimal or desired policy trajectory.

The analytical power of this framework is most apparent when considering the problem of alignment (Christiano, 2018; Russell, 2019). The most salient and complex cases are not perfectly aligned or overtly misaligned agents, but the vast intermediate class of partially aligned systems. For these agents, the paper argues that a quantitative notion of risk becomes tractable. If an optimal or "aligned" policy can be normatively specified, even counterfactually, then risk can be defined and measured as the agent's expected deviation from this baseline. This is made precise by instantiating this paradigm in Markovian agents.

References

Christiano, Paul 2018 "Clarifying "AI alignment"", https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6. (Cit. on p. 1.)

Dung, L 2025 "Understanding AI Agency", Philosophical Studies. (Cit. on p. 1.)

Faroldi, Federico L. G. 2021 "General AI and Transparency in the EU AI Act", i-lex. (Cit. on p. 1.)

—- 2025 "Risk and artificial general intelligence", AI & Society, 40, pp. 2541-2549, doi: 10.1007/s00146-025-02034-y. (Cit. on p. 1.)

Hansson, Sven Ove 2018 "Risk", in The Stanford Encyclopedia of Philosophy, ed. by Edward N. Zalta, Fall 2018, Metaphysics Research Lab, Stanford University. (Cit. on p. 1.)

Kasirzadeh, Atoosa and Iason Gabriel 2025 Characterizing AI Agents for Alignment and Governance, https://arxiv.org/pdf/2504.21848, Manuscript, arXiv: 2504.21848. (Cit. on p. 1.)

Russell, Stuart 2019 Human Compatible, Viking. (Cit. on p. 1.) 2