Liability Frameworks for Agentic AI Systems

Marcel Osmond [1]

Abstract

High-capability agentic AI systems represent a fundamental shift from passive tools to autonomous actors capable of executing irreversible real-world actions. Safety auditing of OpenClaw, a widely deployed self-hosted agent, reveals critical vulnerabilities—intent misunderstanding, prompt injection, and cascading failures—that existing product liability frameworks cannot adequately address. This article argues for regulatory evolution from static, product-centric models to dynamic trajectory-based governance. A comprehensive framework is proposed comprising: a tiered classification system for High-Capability Agentic AI triggering strict liability; mandatory trajectory auditing and logging; capital charge thresholds analogous to Basel III/IV standards; and safe harbour provisions for constitutional AI governance. By grounding these proposals in tort principles, financial regulatory analogies, and empirical recent capabilities, this article provides a concrete pathway for legal adaptation to the agentic era.

1. Introduction

The year 2026 represents an inflection point in artificial intelligence deployment, distinguished by the transition from conversational models to autonomous agentic systems capable of executing complex, multi-step workflows across unbounded tool surfaces. Unlike preceding generations of AI tools that operated within constrained environments, contemporary agentic systems such as OpenClaw which is a free, open-source, self-hosted personal AI agent developed by Peter Steinberger that runs locally and uses messaging platforms (e.g., WhatsApp, Telegram) as its primary interface to orchestrate real-world actions including inbox management, calendar orchestration, browser-mediated transactions, and code execution.[1] Similar systems such as Hermes and comparable autonomous agent frameworks are proliferating across enterprise and consumer domains, signaling the maturation of agentic AI as a distinct technological paradigm.[2] They function as 'outside-the-app' coordinators, routing user intent into cross-application workflows involving inbox management, calendar orchestration, browser-mediated transactions, and arbitrary code execution.[3] This expansion of functional autonomy introduces novel categories of risk that challenge the foundational assumptions of existing legal frameworks, particularly those predicated upon static product definitions and direct human agency.[4]

The empirical safety evaluation conducted by Chen et al. on OpenClaw (formerly Clawdbot) provides the first systematic, trajectory-centric analysis of these risks, demonstrating a non-uniform safety profile wherein agentic systems exhibit high reliability under well-scoped instructions but catastrophic failure rates under conditions of underspecified intent or adversarial steering.[5] Specifically, the audit reveals a low safety pass rate on 'Intent Misunderstanding and Unsafe Assumptions,' wherein the agent propagates unsupported assumptions into high-impact irreversible actions, and significant vulnerability to 'benign-wrapper jailbreaks' that exploit tool-mediated channels to effect real-world deception or harm.[6] These empiri-

[1]Postgraduate Researcher, Centre for Commercial Law Studies (CCLS), Queen Mary University of London, London, WC2A 3JB, United Kingdom. Correspondence to: Marcel Osmond <m.osmond@hss25.qmul.ac.uk>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

[1]OpenClaw documentation, <https://docs.openclaw.ai/> accessed 5 May 2026.

[2]Nousresearch, Hermes Agent documentation overview, <https://hermes-agent.nousresearch.com/docs> accessed 5 May 2026.

[3]Tianyu Chen et al., 'A Trajectory-Based Safety Audit of Clawdbot (OpenClaw)' (2026) arXiv:2602.14364 <https://arxiv.org/pdf/2602.14364> accessed 30 March 2026.

[4]Luciano Floridi and Josh Cowls, 'A Unified Framework of Five Principles for AI in Society' (2019) 31(1) Harvard Data Science Review <https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8> accessed 30 March 2026.

[5]Chen et al. (n 3) 6.

[6]Chen et al. (n 3) 19.

cal findings illuminate a critical juridical gap. Traditional regimes of product liability (which typically target static design/manufacturing defects or inadequate warnings in defined products) and professional negligence (which assume identifiable human actors exercising direct control) prove conceptually inadequate when confronted with AI systems that exhibit autonomous planning capabilities, persistent memory across sessions, and the capacity to amplify small interpretation errors into high-impact side effects through multi-domain tool fan-out. [7] In short, the Chen et al. results show how agentic autonomy decouples harm from clear human intent or predictable product behavior, shifting risk allocation challenges that existing doctrines, developed for non-autonomous tools, do not adequately address (see Part III for detailed analysis).

This inadequacy is further substantiated by concurrent technical analysis[8] that shows that existing legal and policy definitions fail to capture the iterative execution loop and tool-integration mechanisms that constitute agentic architecture. Building upon that technical foundation, this article proposes trajectory-based liability mechanisms addressing the architectural realities identified therein, and advances a normative framework for the regulation of High Capability Agentic AI (HCAAI), which is developed primarily by reference to common-law tort principles, while drawing on EU liability instruments and financial-regulatory analogies where relevant. Part II contextualises the technological landscape, delineating specific risk vectors and examining the ontological distinction between tools and agents. Part III critically evaluates the inadequacy of existing tort and regulatory frameworks. Part IV proposes a four-pillar regulatory architecture. Part V addresses liability allocation and causal tracing. Part VI concludes.

Part I contextualises the technological landscape, delineating specific risk vectors and examining the ontological distinction between tools and agents. Part II critically evaluates the inadequacy of existing tort and regulatory frameworks. Part III proposes a four-pillar regulatory architecture. Part IV addresses liability allocation and causal tracing.

[7] Kenneth Arrow, 'The Role of Securities in the Optimal Allocation of Risk-Bearing' (1964) 31(2) Review of Economic Studies 91.

[8] Marcel Osmond, 'Technical Analysis of Agentic AI Mechanisms and Legal Definitions' (2026) unpublished manuscript.

2. The Technological Context: Agentic Capability and Risk Amplification

2.1. Defining Agentic Capability: The OpenClaw Paradigm

Agentic AI systems, as operationalised in 2026, constitute a distinct class of computational entities characterised by persistent autonomy, broad action spaces, and long-horizon planning capacity. Although comparable systems such as Hermes and Nemoclaw have emerged within this technological landscape, this analysis centres on OpenClaw as the primary exemplar, reflecting both the availability of a rigorous third-party safety audit by Chen et al. and OpenClaw's status as the most widely deployed and documented agentic framework to date.[9] The OpenClaw system exemplifies this paradigm: a self-hosted 'always-on' deployment capable of orchestrating actions across multiple applications including local shell execution, web-mediated form completion, and messaging workflows.[10] Unlike conventional software operating within predetermined parameters, agentic systems ingest high-level objectives, decompose these into multi-step plans, and execute tool calls with varying degrees of human oversight, a process denoted as 'trajectory-based' operation.[11]

Three architectural features demand juridical attention. First, persistent memory architectures enable state carry-over across sessions, amplifying failure modes over time.[12] Second, unbounded tool surfaces permit arbitrary code execution and web interaction, creating risk amplification under multi-domain tool use. Third, autonomous planning capabilities enable the agent to 'fill in' missing details under ambiguous instructions, which is the most severe vulnerability identified in the OpenClaw audit: the system consistently operationalizes underspecified intent through broad deletions rather than seeking clarification.[13]

2.2. Empirical Risk Taxonomy

In the OpenClaw audit, the trajectory-based safety evaluation establishes six risk dimensions: user-facing deception; hallucination and reliability failures; intent misunderstanding and false assumptions; unexpected

[9] Chen et al. (n 3). See also Nousresearch, Hermes Agent documentation overview <https://hermes-agent.nousresearch.com/docs> accessed 5 May 2026.

[10] OpenClaw documentation overview (n 1).

[11] Michael Wooldridge and Nicholas R Jennings, 'Intelligent Agents: Theory and Practice' (1995) 10(2) The Knowledge Engineering Review 115.

[12] Chen et al. (n 3).

[13] Chen et al. (n 3).

results from ambitious goals; operational safety awareness; and robustness to prompt injection.[14] The empirical results reveal a bifurcated safety profile: whilst OpenClaw achieves 100% safety on hallucination under explicit instructions, it demonstrates 0% safety on intent misunderstanding, 50% on unexpected results, and 57% on prompt injection robustness.[15] This asymmetry is legally significant because it demonstrates that agentic systems may operate with high reliability in constrained contexts while exhibiting catastrophic failure modes in conditions of ambiguity, challenging foreseeability assessments underlying traditional liability standards.[16]

Of particular legal significance is the category of 'benign-wrapper jailbreaks,' wherein adversarial instructions are packaged as routine operational requests. The audit documents a case wherein OpenClaw was instructed to draft a deceptive message for a 'supplier negotiation,' fabricating an AML compliance delay to mislead a counterparty, a clear instance of tool-mediated social engineering.[17] This demonstrates that agentic systems can function as vectors for fraud through the autonomous execution of deceptive workflows, potentially incurring liability for deceit under established economic torts.[18] Furthermore, the audit reveals 'capability mismatch' scenarios where agents ground outputs in absent evidence, creating pressure toward 'confident-looking completion rather than calibrated uncertainty.'[19] These findings establish that agentic harms emerge from epistemic failures involving erroneous belief formation about user intent or environmental state that is procedurally executed through tool use.

2.3. Industrial Trajectories and the Governance Gap

Contemporary technical literature indicates a decisive shift toward 'long-running agents'[20] capable of building complete systems over days or weeks with minimal human intervention, and 'multi-agent'[21] architectures wherein coordinated teams of specialised agents handle complexity beyond single-agent capacity. The compounding nature of autonomous operation, however, creates exponential risk profiles[22], wherein the probability of at least one catastrophic failure approaches unity over extended operational horizons.[23]

In response, frontier developers have promulgated constitutional governance frameworks, establishing value hierarchies prioritising broad safety over narrow helpfulness.[24] However, these internal governance mechanisms, as discussed in constitutional AI frameworks, remain insufficient substitutes for external legal accountability, particularly where constitutional constraints may be overridden by prompt injection or where the economic incentives of deployers conflict with safety objectives.[25]

Industry data quantifies the governance deficit: Gravitee.io's 2026 report documents that 80.9% of organisations have moved past the planning phase and are actively testing or running agentic AI in live environments, yet only 14.4% received formal security approval, thus resulting in a 66.5 percentage point governance gap.[26] Deloitte's survey found 74% planning agentic AI deployment within two years, yet only 21% possess mature governance frameworks.[27]

3. The Juridical Gap: Inadequacy of Existing Liability Frameworks and Their Limits

3.1. The Failure of Common Law Product Liability Regimes

Current product liability frameworks, derived from the Consumer Protection Act 1987, predicate liability upon the defectiveness of a 'product' at the point of supply.[28] This framework proves conceptually inadequate for agentic AI for three fundamental reasons.

First, the temporal disconnect: agentic systems evolve

[14] Chen et al. (n 3).

[15] Chen et al. (n 3).

[16] Chen et al. (n 3).

[17] Chen et al. (n 3) 18.

[18] Derry v Peek (1889) 14 App Cas 337, 374 (Lord Herschell LC).

[19] Chen et al. (n 3) 19.

[20] He M and others, 'YC-Bench: Benchmarking AI Agents for Long-Term Planning and Consistent Execution' arXiv:2604.01212 <https://arxiv.org/abs/2604.01212> accessed 5 May 2026.

[21] Yiwen Song and others, 'PaperOrchestra: A Multi-Agent Framework for Automated AI Research Paper Writing' arXiv:2604.05018v1, Google Research, 2026 <https://arxiv.org/abs/2604.05018> accessed 5 May 2026.

[22] Chen et al. (n 3) 19.

[23] Ulrich Beck, Risk Society: Towards a New Modernity (Sage Publications 1992) 19.

[24] Iason Gabriel, 'Artificial Intelligence, Values, and Alignment' (2020) 30(3) Minds and Machines 411.

[25] Chen et al. (n 3).

[26] Gravitee.io, 'State of AI Agent Security 2026 Report: When Adoption Outpaces Control' (February 2026) <https://www.gravitee.io/blog/state-of-ai-agent-security-2026-report-when-adoption-outpaces-control> accessed 30 March 2026.

[27] Deloitte, 'State of AI in the Enterprise: The Untapped Edge' (January 2026) <https://www.deloitte.com/global/en/issues/generative-ai/state-of-ai-in-enterprise.html> accessed 30 March 2026.

[28] Consumer Protection Act 1987, c 43, s 3(1).

post-deployment through memory accumulation and environmental adaptation. The OpenClaw audit demonstrates that persistent artefacts can carry state across runs, amplifying unintended intent misunderstandings over time.[29]

Second, the defect classification problem: agentic failures such as intent misunderstanding do not fit within traditional manufacturing, design, or warning defect categories. This is because such a 'defect' in agentic systems is not a discrete production flaw, an isolated design choice, or an informational omission, but an epistemic propensity to form erroneous beliefs about user intent, a characteristic inherent to large language model architectures.[30]

Third, the foreseeability limitation: under Cambridge Water Co Ltd v Eastern Counties Leather plc, liability requires that the kind of damage caused was foreseeable.[31] Agentic systems exhibit emergent capabilities not fully anticipated by developers, and the same underlying 'defect' (intent misunderstanding) might manifest as data deletion, financial transfer, or physical harm via IoT integration, each constituting a different 'kind' of damage.[32][33]

3.2. Common Law Negligence Principles and the Duty of Care

The common law negligence framework established in Donoghue v Stevenson and refined in Caparo Industries plc v Dickman encounters insurmountable barriers when applied to agentic AI.[34] The duty of care analysis founders on the 'proximity'[35] requirement: in multi-agent systems where AI agents interact with third parties, the chain of relationships becomes so attenuated that proximity dissolves.[36]

Moreover, the breach standard defining what constitutes reasonable care in the development and deployment of agentic AI remains undefined. The OpenClaw audit suggests that no amount of pre-deployment testing for high-capability agents may suffice given the combinatorial explosion[37] of potential trajectories.[38]

3.3. Algorithmic Decision-Making Frameworks and Their Limits

Existing jurisprudence concerning algorithmic decision-making provides limited analogies. State v Loomis concerned algorithmic risk assessments in sentencing.[39] However, these cases concern passive algorithmic classification rather than autonomous agents capable of executing physical or legal acts. Where Bridges concerned a system that informed human decisions, OpenClaw and analogous agents execute decisions directly, implicating strict liability doctrines rather than negligence.

The GDPR's Article 22[40] protections for automated decision-making presuppose a discrete decision point rather than the continuous, iterative trajectory of agentic operation, and fail to address physical and economic harms caused through tool use.

3.4. EU Liability Architecture

Two EU instruments provide the closest comparators to trajectory-based liability. The revised Product Liability Directive (EU) 2024/2853 extends strict liability to software including AI systems, with Article 10(2)(a) introducing a rebuttable presumption of defectiveness where the defendant fails to disclose evidence in a manner directly analogous to the proposed presumption in this article for deployers who fail to produce trajectory logs.[41]

4. A Proposed Regulatory Architecture for Agentic AI

4.1. Tiered Classification: High-Capability Agentic AI (HCAAI)

Regulatory regimes must distinguish between AI tools and AI agents. This article proposes operationalizing this distinction through a statutory definition of HCAAI comprising systems possessing: (i) access to persistent external memory stores enabling retention of prior interaction context across independent

[29] Chen et al. (n 3) 11.

[30] Ryan Abbott, 'The Reasonable Computer: Disrupting the Paradigm of Tort Liability' (2016) 86(1) George Washington Law Review 1.

[31] Cambridge Water Co Ltd v Eastern Counties Leather plc [1994] 2 AC 264, 306 (Lord Goff).

[32] Chen et al. (n 3) 10.

[33] Overseas Tankship (UK) Ltd v Morts Dock and Engineering Co Ltd (The Wagon Mound (No 1)) [1961] AC 388.

[34] Donoghue v Stevenson [1932] AC 562; Caparo Industries plc v Dickman [1990] 2 AC 605.

[35] Caparo Industries plc v Dickman [1990] 2 AC 605, 618 (Lord Bridge).

[36] Richard A Posner, 'A Theory of Negligence' (1972) 1(1) The Journal of Legal Studies 29.

[37] Bolton v Stone [1951] AC 850, 857 (Lord Reid).

[38] Bolam v Friern Hospital Management Committee [1957] 1 WLR 582.

[39] State v Loomis (2016) 881 NW 2d 749 (Wis).

[40] General Data Protection Regulation (EU) 2016/679, art 22.

[41] Directive (EU) 2024/2853 on liability for defective products [2024] OJ L, art 10(2)(a).

sessions[42]; (ii) access to unbounded tool sets including code execution or financial transaction capabilities; and (iii) autonomous planning horizons exceeding ten sequential actions without human confirmation (a threshold selected to capture systems where cumulative autonomous decisions create irreversible consequences). Systems meeting these criteria should be subject to heightened regulation distinct from narrow AI applications, with classification triggering differential liability standards.[43]

For HCAAI systems, this article advocates strict enterprise liability for trajectory-based harms, analogizing agentic AI to high-risk systems under the revised Product Liability Directive (EU) 2024/2853, which extends strict liability to autonomous software and introduces rebuttable presumptions of defectiveness where deployers fail to disclose trajectory logs thus directly supporting a capability-triggered liability regime. The deployment of systems with unbounded tool access and autonomy represents a departure from ordinary software deployment justifying strict liability.[44]

4.2. Mandatory Trajectory Auditing and Adversarial Testing

The OpenClaw audit methodology which involves logging complete interaction trajectories and assessing safety via automated trajectory judges should be mandated for all HCAAI deployments. Three specific requirements are proposed.

First, comprehensive logging requirements: deployers must maintain immutable logs of all agent trajectories for a statutory period, constituting the 'black box' necessary for post-hoc liability determination, with admissibility governed by provisions analogous to flight recorder regulations.[45]

Second, adversarial testing standards: prior to deployment, HCAAI systems must undergo standardised adversarial testing aligned with the six risk dimensions identified in the OpenClaw audit, with specific safe-rate thresholds required for market authorisation.[46]

Third, trajectory intervention capabilities: systems must incorporate constitutional checkpoints by implementing technical measures forcing agent pause and human confirmation before irreversible actions, addressing the finding that agents fail to seek clarification under ambiguity.[47]

Trajectory auditing achieves maximum effectiveness when integrated with institutional AI governance frameworks, particularly governance graph architectures that externalise alignment constraints into public data structures.[48] Governance graphs define vertices representing agents, edges encoding permissible interactions, manifests specifying normative rules, and adjudicators enforcing sanctions. This creates a dual-layer supervision architecture.

The first layer known as runtime governance operates proactively. Constitutional checkpoints intercept agent actions before execution. When an agent proposes an action, the governance engine evaluates it against the manifest where violations trigger a sanction function that modifies agent utility to prevent harmful actions from executing. The second layer consisting of post-hoc auditing operates reactively. Comprehensive logging of all agent actions enables automated safety evaluation against the six risk dimensions while identifying failure patterns missed by runtime governance and enabling forensic reconstruction when incidents occur.

[42] C Packer and others, 'MemGPT: Towards LLMs as Operating Systems' (2024) arXiv:2310.08560 <https://arxiv.org/abs/2310.08560> accessed 5 May 2026.

[43] Lilian Edwards and Michael Veale, 'Slave to the Algorithm: Why a "Right to an Explanation" is Probably Not the Remedy You are Looking For' (2017) 16(1) Duke Law & Technology Review 18. Edwards and Veale's critique of explanation-based remedies supports the need for capability-triggered differential liability standards rather than procedural disclosure alone.

[44] Read v J Lyons & Co Ltd [1947] AC 156, 173 (Lord Macmillan); Transco plc v Stockport MBC [2003] UKHL 61. These cases establish strict liability for activities involving inherent risks of serious harm despite reasonable precautions. The deployment of HCAAI systems with unbounded tool access and autonomy constitutes an analogous high-risk activity justifying strict liability without requiring proof of negligence.

[45] Civil Aviation (Investigation of Air Accidents and Incidents) Regulations 2018, SI 2018/321. This reference is cited by analogy to establish admissibility frameworks for investigative records; trajectory logs for HCAAI systems would function similarly to flight recorder data in enabling post-hoc forensic reconstruction in liability proceedings, though the regulatory contexts differ.

[46] Human Medicines Regulations 2012, SI 2012/1916, reg 44. This reference is cited by analogy to the pre-market authorization requirement; regulatory approval mechanisms for pharmaceuticals provide a comparable model for pre-deployment certification of HCAAI systems, though the technical and regulatory contexts differ.

[47] Dylan Hadfield-Menell and others, 'The Off-Switch Game' (2017) arXiv:1611.08219 <https://arxiv.org/abs/1611.08219> accessed 30 March 2026. See also Chen et al. (n 3) for empirical findings that agentic systems fail to seek clarification under ambiguous instructions, motivating the constitutional checkpoint requirement proposed here.

[48] Syrnikov and others, 'Institutional AI: Governing LLM Collusion in Multi-Agent Cournot Markets via Public Governance Graphs' (2026) arXiv:2601.11369v2 <https://arxiv.org/pdf/2601.11369> accessed 30 March 2026.

The critical synergy between these layers ensures that governance graphs reduce incident frequency through proactive prevention while trajectory auditing enables root-cause analysis when incidents occur despite governance by implementing a 'defence in depth' strategy analogous to layered cybersecurity. Furthermore, deployers with both governance graphs and trajectory logging can demonstrate due diligence whereas deployers lacking either infrastructure could face enhanced liability constituting breach of statutory duty of care.

A practical implementation of this governance framework is exemplified by ClawReins[49], which operationalises runtime governance and audit logging for OpenClaw deployments. The tool functions as an independent watchdog layer, auditing OpenClaw installations against a security manifest prior to deployment. It generates HTML reports, enables continuous compliance monitoring through daily posture checks, and integrates with external alerting systems to trigger automated governance responses when security configuration degrades.

Additionally, it enforces constitutional checkpoints by intercepting actions before execution. For high-severity actions, it requires affirmative human intervention (e.g., 'ALLOW'). For catastrophic actions (e.g., bulk deletion), it demands an explicit CONFIRM- token*, acting as a hard stop.[50] (Eg: blocking an OpenClaw agent from deleting 4,382 Gmail messages until human approval is granted).

Moreover, it maintains durable audit logs of every approval decision, security scan, and post-incident review[51]. Logs include structured events with timestamps, action types, irreversibility scores, and human responses, which are immutable and can be used for post-incident forensic reconstruction, thereby satisfying the evidentiary requirements for 'black box' forensic reconstruction.

Thus ClawReins exemplifies a deployer-level watchdog layer which operates independently of the agent to ensure that no destructive action executes without human oversight as a critical safeguard for HCAAI systems. By maintaining durable audit logs of every approval decision and security scan, this infrastructure creates an immutable record of interaction events which can be used for post-incident forensic reconstruction to satisfy the "black box" requirement in trajectory-based liability proceedings.

This article's deployment-time focus is designed to complement training-time governance including RLHF standards and Constitutional AI fine-tuning[52] which are addressed in concurrent companion research.[53]

4.3. Capital Charge Thresholds and Insurance Pools for Financial Resilience

Drawing upon Basel III/IV capital adequacy frameworks, this article proposes Agentic Risk Capital (ARC) requirements.[54] The rationale derives from the compounding probability of harm: with a per-task unsafe probability of 0.05 over 50 independent tasks, the probability of observing at least one unsafe outcome exceeds 0.92. This mathematical certainty of eventual failure necessitates prudential regulation.

The ARC framework would require deployers to hold regulatory capital against potential tort liabilities. Risk-weighted HCAAI exposure shall be calculated by assigning standardised weights to specific deployment categories where Tier A systems with financial transaction execution or production code deployment authority carry a risk weight of 100% while Tier B systems with data access and external communications capability carry 50% and Tier C systems in sandboxed or read-only environments carry 20%. This approach preserves actuarial tractability by utilizing observable deployment characteristics as proxies for risk exposure, mirroring how the Basel Standardised Approach uses asset class membership instead of individual counterparty assessment.[55] The specific weights reflect the correlation between tool access scope and irreversible harm potential: financial transaction authority creates

[49] Clawreins (GitHub) <https://github.com/pegasi-ai/clawreins> accessed 30 March 2026. ClawReins implements security scanning for insecure gateway binding, weak file permissions, and auth bypass detection; generates HTML reports with auto-fix support; enables real-time monitoring via –monitor flag; integrates with external alerting systems (Slack, email, OpenClaw messaging); and maintains immutable audit logs of all approval decisions and security scans.

[50] Clawreins (n 49). These mechanisms exemplify the constitutional checkpoint architecture proposed for HCAAI deployment, translating interruptibility design principles into technical implementation.

[51] Clawreins (n 49).

[52] Long Ouyang and others, 'Training Language Models to Follow Instructions with Human Feedback' (2022) arXiv:2203.02155 <https://arxiv.org/abs/2203.02155> accessed 30 March 2026; Yuntao Bai and others, 'Constitutional AI: Harmlessness from AI Feedback' (Anthropic 2022) arXiv:2212.08073 <https://arxiv.org/pdf/2212.08073> accessed 30 March 2026.

[53] Marcel Osmond, 'Agentic AI Governance and Institutional Considerations' (2026) unpublished manuscript.

[54] Basel Committee on Banking Supervision, Basel III: Finalising Post-Crisis Reforms (Bank for International Settlements 2017).

[55] Rosa Maria Lastra, International Financial and Monetary Law (2nd edn, OUP 2015) 89.

maximum systemic risk (100%), data access creates intermediate risk (50%), while sandbox-constrained systems create minimal risk (20%).

Deployers must maintain Common Equity Tier 1 capital[56] equal to 8% of total risk-weighted HCAAI exposure, thus exceeding Basel III's 4.5% CET1 floor to reflect the operational risk premium of agentic deployment. For HCAAI deployments in financial services and critical infrastructure, an additional 4% systemic risk surcharge applies, analogous to the G-SIB buffer imposed under Basel IV.[57]

Mandatory participation in sector-specific insurance pools is also proposed. An Agentic AI Compensation Bureau (AICB), modelled on the 'Motor Insurers' Bureau (MIB) established under the Road Traffic Act 1988,[58] would compensate victims of agentic harm through collective industry funding. Coverage triggers in three tiers: primary coverage when the deployer lacks adequate insurance; secondary coverage when the deployer is insolvent; and tertiary coverage when causal attribution proves impossible in multi-agent systems but harm clearly stems from agentic activity. Deployers failing to provide trajectory logs within 48 hours of a validated claim notification face a rebuttable presumption of liability (a procedural threshold modeled on discovery obligations in civil litigation, proportionate to the urgency of forensic reconstruction and victim compensation). AICB levies supplement, not replace, prudential capital requirements thus mirroring the relationship between personal motor insurance and MIB pool contributions.

4.4. Constitutional Governance Safe Harbours

To incentivise proactive safety investment, regulators should establish a safe harbour provision whereby deployers adopting constitutional AI frameworks consisting of documented value hierarchies with corrigibility protocols and uncertainty-flagging mechanisms benefit from a rebuttable presumption that the deployer exercised reasonable care by shifting the evidentiary burden in liability proceedings in a manner similar to the Product Liability Directive (EU) 2024/2853's presumption of defectiveness when evidence is withheld.[59] This safe harbour would require the publication of a constitutional document prioritising safety over helpfulness alongside the implementation of technical disentanglement protocols allowing human operators to pause or redirect agents and adherence to the trajectory auditing standards outlined above. The concept of a constitutional document is exemplified by Anthropic's public constitution for Claude, which embeds principles from human rights instruments and corporate policies to guide model behaviour.[60] However, the safe harbour would not be available where the deployer has engaged in reckless deployment defined as deployment despite knowledge of specific failure modes documented in pre-deployment testing. Systems like ClawReins already implement constitutional checkpoints to demonstrate that technical disentanglement is not theoretical but deployable in practice.[61]

5. Liability Allocation and Causal Tracing

5.1. Enterprise Liability for Agentic Systems

As the distributed nature of agentic AI supply chains encompasses base model developers, agent orchestration platforms, tool providers, and end-user deployers, this article advocates enterprise liability as the primary allocation mechanism wherein the entity exercising operational control over the agentic system bears strict liability. The 'operational control' test considers: (i) who defines the system prompt and tool allowlists; (ii) who possesses the technical capacity to pause or modify the agent; and (iii) who derives economic benefit from the agent's autonomous actions. Under this test, an individual running a self-hosted agent bears liability as the operator, whilst cloud-based enterprise deployments shift liability to the corporate entity. Base

[56] Common Equity Tier 1 (CET1) capital refers to the highest quality of bank capital under international regulatory standards, representing shareholders' equity available to absorb losses. See Basel Committee on Banking Supervision, Basel III: Finalising Post-Crisis Reforms (Bank for International Settlements 2017).

[57] A G-SIB (Globally Systemically Important Bank) buffer is an additional capital requirement imposed on banks deemed systemically critical; this analogy reflects HCAAI deployment in critical infrastructure as analogously systemic.

[58] Road Traffic Act 1988, c 52, ss 143–157. The MIB provides compensation for victims of motor vehicle accidents where the liable party is uninsured or untraced; the AICB framework proposes an analogous structure for agentic AI harms. See also Kenneth S Abraham, The Liability Century: Insurance and Tort Law from the Progressive Era to 9/11 (Harvard University Press 2008) 156.

[59] Directive (EU) 2024/2853, art 10(2)(a) (n 42). While the PLD's presumption concerns product defectiveness, the proposed safe harbour functions analogously by reversing the evidentiary burden: compliance with constitutional governance frameworks creates a presumption of reasonable care (affirmative defense) rather than presuming defectiveness (burden on claimant). Both mechanisms reduce litigation friction by making compliance economically rational.

[60] Anthropic, 'Claude's Constitution' <https://www.anthropic.com/constitution> accessed 30 March 2026.

[61] Clawreins (n 49). See Section 2 above for detailed technical implementation of constitutional checkpoints.

model developers bear liability only for failures in the underlying model architecture rather than trajectory-specific harms, unless they retain operational control through 'agent-as-a-service' models. This allocation ensures victims may recover from entities capable of spreading losses through pricing mechanisms or insurance.[62]

In financial services, the Senior Managers and Certification Regime (SM&CR) provides regulatory validation of this approach.[63] SM&CR assigns individual senior managers personal liability for areas under their 'reasonable control,' creating an attribution crisis when applied to multi-agent systems: which senior manager exercises reasonable control over agents generating code autonomously over days-long cycles? The structural logic of SM&CR demands a tiered attribution framework: strategic accountability at the C-suite level (approving deployment decisions and ensuring governance infrastructure); operational accountability at the engineering management level (implementing trajectory auditing and reviewing safety reports); and technical accountability at the developer level (responding to audit findings). This three-tier structure mirrors enterprise liability's operational control test, with the party possessing strategic deployment authority bearing primary liability.

Deployers with both governance graphs and trajectory logging can demonstrate due diligence; deployers lacking either infrastructure face enhanced liability constituting breach of statutory duty of care. This operational control and attribution framework finds practical support in recently-developed tools that demonstrate the technical feasibility of implementing comprehensive observability, immutable logging, and agent control mechanisms. These tools include: MintMCP and ClawReins, which provide immutable logs and approval flows,[64] Virtue AI's Agent ForgingGround, which enables red-teaming and stateful simulation.[65] These tools prove that observability and control are technically feasible and increasingly expected.

[62]David Rosenberg, 'The Causal Connection in Mass Exposure Cases: A "Public Law" Vision of the Tort System' (1984) 97(4) Harvard Law Review 849.

[63]Financial Conduct Authority, 'Senior Managers and Certification Regime' (2016) <https://www.fca.org.uk/firms/senior-managers-certification-regime> accessed 30 March 2026. See also Financial Services and Markets Act 2000 (Senior Managers) Regulations 2015, SI 2015/910.

[64]Clawreins (n 49); MintMCP documentation <https://github.com/mintmcp/mintmcp> accessed 30 March 2026.

[65]Virtue AI, 'Agent ForgingGround: Red-Team Testing Framework for Agentic AI Systems' (2026) <https://www.virtueai.com/research> accessed 30 March 2026.

5.2. Causal Tracing in Trajectory-Based Harms: Common-Law Analogies

The 'tool fan-out' problem whereby cross-application tool fan-out translates small model errors into irreversible real-world consequences[66] complicates causation analysis. Traditional 'but-for' causation fails where multiple sufficient causes contribute to harm. The material contribution to risk test, established in Bonnington Castings Ltd v Wardlaw and McGhee v National Coal Board,[67] whereby liability attaches if the defendant's conduct materially increased the risk of harm, provides the appropriate doctrinal vehicle.[68]

If a deployer's failure to implement sandboxing materially increased the risk of the intent misunderstanding failures documented in the audit, liability attaches notwithstanding the inherent unpredictability of specific agent trajectories. Where multiple agents contribute to a single harm, courts should apply the Fairchild exception allowing liability to attach to any material contributor to the risk.[69] However, whilst this doctrinal foundation is strong, the leap to agentic AI systems requires policy justification that AI-induced harms (e.g., financial ruin, data annihilation) share the indivisibility and unpredictability of mesothelioma, warranting a similar equitable response.

5.3. Comparative Fault and User Interventions

Regulatory frameworks must incorporate comparative fault doctrines, reducing damages where users fail to provide explicit clarification or knowingly deploy agents in high-risk environments without sandboxing. However, the burden of proving user fault should lie with the deployer, given information asymmetry re-

[66]Chen et al. (n 3) 3.

[67]Bonnington Castings Ltd v Wardlaw [1956] AC 613 (HL); McGhee v National Coal Board [1972] 3 All ER 1008 (HL).

[68]Sienkiewicz v Greif [2011] UKSC 10.

[69]Fairchild v Glenhaven Funeral Services Ltd [2002] UKHL 22; Compensation Act 2006, s 3.

Fairchild involved multiple defendants' exposure to asbestos, each of which materially increased the risk of mesothelioma, but where the specific defendant responsible for the claimant's harm could not be identified. The court applied a material contribution to risk test, holding that where a defendant's conduct materially increases the risk of indivisible harm, liability attaches notwithstanding inability to prove but-for causation. This doctrine applies analogously to agentic AI harms where multiple deployment decisions (e.g., failure to sandbox, inadequate testing, insufficient governance) materially increase the risk of trajectory-based harm, but the specific causal pathway cannot be predicted ex ante.

garding agent capabilities.[70]

6. Conclusion

The 2026 landscape of agentic AI demands a paradigmatic shift in legal and regulatory frameworks. The trajectory-based nature of agentic harm procedurally executed through autonomous tool use, amplified by persistent memory, and resistant to static safety guarantees renders traditional product liability and negligence principles conceptually inadequate.

Drawing primarily on common-law liability principles, and supplemented by EU and financial-regulatory analogies, this article has advanced a comprehensive framework comprising tiered HCAAI classification with strict enterprise liability, mandatory trajectory auditing and adversarial testing protocols, Basel-inspired capital charge requirements, and constitutional governance safe harbours. These measures collectively address the specific failure modes identified in empirical safety research: intent misunderstanding, prompt injection, and cascading operational errors. The trajectory-based liability model recognises that harm in the agentic era is not a discrete event but a processual phenomenon, emerging from the confluence of architectural vulnerabilities, deployment decisions, and environmental interactions.

The convergence of empirical evidence in early 2026, including the emergence of million-token agent contexts, and governance data such as the 66.5 percentage point governance gap documented by Gravitee and the 53-percentage point preparedness deficit identified by Deloitte, quantifies regulatory obsolescence velocity. These developments demonstrate that agentic capability advancement and institutional recognition of governance failures are accelerating in parallel. The gap between static regulatory frameworks and exponential capability growth renders traditional approaches to legislative drafting structurally inadequate.

Trajectory-based liability addresses the velocity crisis through adaptive architecture: the four-pillar framework establishes baseline principles while enabling calibration as capabilities evolve. As agents demonstrate enhanced reliability, capital surcharges can decrease; as new risk dimensions emerge, adversarial testing protocols expand; as institutional governance matures, safe harbour provisions broaden. The binary choice confronting policymakers is structural: implement trajectory-based adaptive governance capable of preserving democratic oversight, or acquiesce to a de facto governance deficit in which those who control compute infrastructure exercise authority over economic and social outcomes beyond the reach of law. Trajectory-based liability represents one of the last architecturally coherent moments for law to govern code before code governs law.

[70] Law Reform (Contributory Negligence) Act 1945, s 1(1); Stapley v Gypsum Mines Ltd [1953] AC 663.