Stability-Assured Framework for Entities (SAFE)

An Agent-Agnostic Risk Assessment Framework for Agentic Systems: A Control Theory Approach

Dr. David R. Hardoon
www.davidroihardoon.com

25 November 2025
Version 4.0

A Personal Note: From FEAT to SAFE – Simplifying AI for All

I've been driven by a singular conviction ever since my early involvement in shaping the Monetary Authority of Singapore's (MAS) FEAT Principles in 2018: AI must be demystified, not deified. FEAT—Fairness, Ethics, Accountability, and Transparency—was never about erecting barriers but about translating AI's complexities into the familiar lexicon of everyday operations and ethical imperatives. By grounding AI in principles that resonate with existing governance practices, we sought to normalize it as a tool of the people, accessible across organizations, not confined to elite silos or fetishized as an exotic novelty.

This paper introduces the Stability-Assured Framework for Entities (SAFE) — previously circulated in draft form as the Agent-Agnostic Risk Assessment Framework (ARAF). The new name reflects the framework's ultimate promise: delivering proven, quantifiable stability to any decision-making entity in a closed-loop system, whether human, rule-based, or artificial.

This ethos endures as agentic AI emerges in 2025, promising autonomy in everything from financial trading to healthcare coordination. Yet, there's a peril in treating "agents" as a radical departure, demanding bespoke mindsets and labyrinthine frameworks. I hold firm: true progress lies in the opposite—forcing simplicity to reveal where formulas must evolve. SAFE embodies this, drawing on control theory's legacy of taming dynamic systems, from aviation to power grids. By viewing every agent—human, rule, or LLM—as a controller in a feedback loop, SAFE strips away the hype, offering a universal, quantifiable language for safe autonomy.

My hope? That SAFE sparks a renaissance where AI becomes as unremarkable as electricity: ubiquitous, reliable, and empowering. Boards and regulators, armed with these metrics, can foster innovation without fear, ensuring agentic systems serve humanity's grand designs. In simplifying the profound, we unlock a future where AI elevates us all—not as a distant dream, but as a daily ally.

Dr. David R. Hardoon

Table of Contents

Foreword to Regulators: Leveraging SAFE for Robust AI Governance

As regulators encounter the surge of agentic AI in 2025—from autonomous financial agents to multi-agent healthcare systems—a unified, quantitative framework for systemic risks is essential. Traditional tools may falter against dynamic feedback loops that can escalate disturbances into market herding or grid failures, as BIS 2024–2025 reports warn.

The Agent-Agnostic Risk Assessment Framework (SAFE) provides a control-theoretic lens, agnostic to controllers (human, rule, or LLM), evaluating risks via observability, controllability, stability, robustness, and performance. Regulators can harness SAFE to:

- **Standardize Compliance**: Mandate loop scoring (1–5) for audits, aligning with EU AI Act tiers, MAS November 2025 Guidelines, and NIST AI RMF—shifting from checklists to metrics that curb arbitrage.
- **Probe Systemic Threats**: Apply robustness tests (e.g., Monte-Carlo herding simulations) to set macro-prudential limits on agent concentration, echoing BIS on positive feedbacks.
- **Enable Safe Scaling**: Define autonomy levels (0–5) for sandboxes, bounding risks while accelerating innovation, complementing long-term alignment research.

Rooted in decades of safety engineering from aviation to nuclear, SAFE equips regulators to govern AI as controllable infrastructure.

Foreword to Board Directors: Harnessing SAFE for Strategic AI Oversight

Boards face a pivotal moment in 2025: agentic AI promises efficiency gains in trading and supply chains but risks cascades from unchecked loops, per BIS alerts on flash crashes. Regulators increasingly expect directors to exercise vigilant, enterprise-wide oversight of AI initiatives and attendant risks, ensuring accountability permeates every layer.

SAFE delivers a boardroom metric—control-theoretic scoring of loop properties—to quantify safe autonomy, treating AI like human or legacy systems. Directors can use it to:

- **Set Fiduciary Limits**: Benchmark scores against ISO 42001 and MAS FEAT, enforcing delegation thresholds that protect value.
- **Guard Enterprise Resilience**: Evaluate multi-agent hierarchies for instability, pre-empting concentration risks in volatile markets.
- **Fuel Assured Growth**: Mirror aviation standards to audit AI as a core asset, unlocking scale without tail-risk exposure.

SAFE turns governance into strategy, enabling confident AI adoption akin to nuclear's regulated boom. Lead with metrics; secure tomorrow's edge.

1. Executive Summary

The rapid emergence of truly agentic AI systems has exposed the limitations of existing AI risk and governance frameworks. These frameworks were designed for an era when AI was primarily a passive prediction tool. Today, real-world deployments — from OpenAI's ChatGPT Atlas and Perplexity's Comet agentic browsers (launched October 2025) to enterprise multi-agent platforms — show agents that autonomously browse, read, decide and act across a user's entire digital life.

Many practitioners still view "agentic AI" narrowly as "AI that performs actions". This premise is dangerously limited. The true locus of risk lies not in the controller's form (human, rule-based, model, or LLM) but in the closed feedback loop it forms with the real world. A framework unfit to assess a human trader, legacy COBOL script, or frontier LLM agent equally fails at scale. Regulators now demand vigilant, enterprise-wide oversight from boards, while practitioners risk overcomplicating "agentic" deployments with bespoke mindsets, echoing the hype that once confined AI to silos.

The proposed Agent-Agnostic Risk Assessment Framework (SAFE) assesses risk as an emergent property of the loop itself—via five timeless metrics: observability (measuring internal states), controllability (override capabilities), stability (equilibrium recovery), robustness (tolerance to disturbances), and performance (alignment with references). By modelling every agent as a controller interacting with an environment (the real-world process), SAFE enforces agent-agnostic evaluation, hierarchical supervision for multi-agent fleets, and bounded autonomy levels (0–5), ensuring validity across capability regimes.

By giving boards and regulators the same rigorous safety engineering language already used for aviation, nuclear plants and power grids for over 80 years, SAFE removes the final obstacle to broad, confident commercial adoption of agentic AI. Though contextualized for finance, SAFE transcends domains, from healthcare coordinators to supply-chain optimizers. It complements static frameworks, accelerates compliant innovation, and paves the way for AI's maturation into ubiquitous infrastructure. By forcing simplicity, SAFE reveals where paradigms must adapt, unlocking a future where agentic systems empower without peril: measurable, governable, and profoundly human-aligned.

2. Introduction

Traditional AI governance frameworks, from the OECD AI Principles and UNESCO recommendations to ISO/IEC 42001:2023, NIST AI RMF, EU AI Act, South Korea's Framework Act on AI (2025), Japan's AI Guidelines for Business v1.1 (2025), China's Artificial Intelligence Safety Governance Framework 2.0 (Sep 2025), and Singapore's MAS FEAT and November 2025 AI Guidelines, remain overwhelmingly focused on static risks: bias, explainability, data quality and model-level accountability.

These concerns remain important, but become secondary the moment systems become agentic. An agent that can browse, email, trade, modify prompts or spawn sub-agents introduces risks that are fundamentally dynamic and systemic: a perfectly "fair" model can still trigger flash crashes, bidding wars, apology loops or planetary-scale positive-feedback cascades.

The Bank for International Settlements has issued the clearest warnings on these risks in its 2024–2025 papers, explicitly calling out positive-feedback loops, herding, concentration risk and the destabilising potential of AI agents.

Control theory has been solving exactly these dynamic feedback risks for over eight decades in every safety-critical domain. Observability, controllability, stability margins, robustness and hierarchical control are technology-agnostic concepts that have scaled to enormous complexity. Recent work has begun applying control-theoretic ideas directly to LLM-based agents (Guo et al., 2024; Zhang et al., 2025; Jia et al., 2025; Zahedifar et al., 2025), confirming the timeliness of the approach.

Anthropic's December 2024 engineering note similarly observes that the most successful agentic implementations overwhelmingly favour simple, composable patterns — precisely what good control engineering demands.

Governance is innovation. It must be acknowledged as such and brought to life through simplification and standards, not through ever-growing checklists. By grounding agentic risk management in control theory, we finally give regulators, boards and engineering teams a shared, quantitative language for "how much autonomy is safe?", regardless of whether the controller is human or machine.

This controller-agnostic property is not academic elegance; it is an absolute requirement for the hybrid human–AI systems that already dominate 2025 deployments and for the severe emerging risks that are classic control problems in disguise.

3. Agent-Agnostic Risk Assessment Framework (SAFE)

Core Model – Every Agent is a Controller

The system receives a Desired outcome, called the Reference input. The closed loop then operates as follows:

Controller (human, rule, model or AI agent) receives the Reference and the feedback Observations → forms a Decision → Actuator translates the Decision into Action → Action affects the Environment[1] (the real-world process) → Sensors capture Observations from the Environment and feed them back to the Controller, closing the loop. Disturbances can enter at any point (market shocks, adversarial inputs, data drift, human error).

Some may object that a sufficiently advanced AI agent could operate without any reference at all. This objection does not survive contact with reality: true agency without directed purpose collapses into randomness, not coherent behaviour; the very act of deployment imposes a reference (explicit or implicit) through organisational intent, training objectives, or the minimum requirement to be useful without harm; even today's most open-ended frontier models carry strong implicit references shaped by human values during training; control theory has long ago discarded open-loop designs as unstable and unfit for any consequential application; and no regulator will permit a high-impact system whose designers claim "it has no goal". A professed reference-free agent is therefore either undeployable noise or an abdication of responsibility. The reference always exists; the only choice is whether we acknowledge it and engineer the loop accordingly.

The framework evaluates risk via five control-theoretic properties:

1. **Observability**

   Can we adequately measure the relevant internal states and external actions in real or near-real time, ensuring data provenance and reliability (e.g., no unverified inputs from untrusted sources)? Mitigation strategies: full Chain-of-Thought (CoT) logging (capturing step-by-step reasoning and action traces), decision provenance, shadow human review, periodic state snapshots, and input validation filters (e.g., source authentication for tools/APIs).

2. **Controllability**

   Can we override, rate-limit, or redirect the controller when needed, including identity-bound enforcement? Mitigation strategies: hierarchical supervision, hard constraints, circuit-breakers, human approval gates, and role-based access controls (e.g., MFA-linked overrides).

3. **Stability**

   Does the loop return to equilibrium after disturbance, or diverge/oscillate? Assessment: red-teaming, historical shock replay, loop-gain analysis.

[1] In Control-Theory the term "Plant" refers to the real-world process or system being controlled and observed—is standard (originating from chemical engineering contexts like factories). For broader accessibility, especially in AI, governance, or non-engineering audiences, we use the more intuitive alternative "Environment".

4. **Robustness**

   How much uncertainty (prompt drift, tool failure, adversarial input, model updates) can the loop tolerate? Measured via gain/phase margins or Monte-Carlo stress testing.

5. **Performance**

   Settling time, overshoot, steady-state error against reference.

These can be scored 1–5 (1 = critically deficient / high risk → 5 = excellent / low risk) or qualitatively as Low/Medium/High, then aggregated into an overall Loop Risk Profile that directly determines the maximum safe Autonomy Level (0–5). Mitigations always increase the score; the goal is to raise it to the institution's chosen threshold.

Important note on applicability: Classical control theory often assumes linear, time-invariant environments for analytical tractability. Neither humans nor LLMs satisfy these assumptions — both are highly non-linear, time-variant, and high-dimensional. Control engineering has successfully managed precisely such non-linear, time-variant systems for decades (e.g., human-in-the-loop aviation, adaptive chemical process control) using exactly the extensions SAFE employs: robust control, gain scheduling, adaptive control, and hierarchical decomposition.

Hierarchical & Multi-Agent Extension (Why It Matters and How It Works)

Most real-world systems are not single loops but hierarchies of loops, exactly as modern organisations already are: junior staff report to seniors, teams report to department heads, and the board oversees the CEO. In agentic systems the same structure emerges naturally and is essential for safety at scale.

A supervisory controller (Level n) issues References to one or more subordinate controllers (Level n−1), receives their Observations, and retains the ability to override or constrain them. This creates clean separation of concerns, bounded autonomy, and clear accountability chains, and — most importantly — the ability to detect and interrupt positive-feedback loops before they become systemic.

Classic financial example: four procurement agents negotiating independently can create a bidding war (positive feedback, loop gain > 1). A supervisory agent with visibility of total budget immediately sees the collective deviation, applies damping (rate limits, consensus rules, or random delays), and restores stability. The same pattern applies to trading desks, fraud-detection swarms, or customer-service agent teams that risk apology/escalation loops.

For maximum robustness, supervisors should where possible employ diversity (different model families, versions, or even non-LLM overseers) to reduce common-mode failure risk. Hierarchies are not bureaucracy here; they are the proven engineering solution for achieving both high performance and high safety in complex systems.

4. How to Use the Framework in Practice

Step-by-Step Process

1. Map the control loop(s) – identify Reference, Controller(s), Actuators, Environment, Sensors
2. Score the five properties (use scorecard)
3. Calculate Overall Score → Autonomy Level
4. Apply mitigations until threshold reached
5. Issue one-page SAFE Assurance Statement + review schedule

Making Scores Objective (Avoiding Subjectivity Traps)

The most common criticism of any scorecard is "who decides the number?". The SAFE can avoid this by tying scores to measurable proxies that institutions can implement, for example:

- **Observability** → % of decisions where full reasoning trace is reconstructible within 5 seconds (>95 % = 5; <50 % = 1)
- **Controllability** → Measured override latency (99th percentile <100 ms = 5; >10 s = 1) + number of independent privilege boundaries
- **Stability** → Worst-case overshoot observed in last 100 red-team or historical scenarios (<10 % = 5; >100 % = 1)
- **Robustness** → Maximum adversarial prompt/edit distance survived in standardised red-team suite (e.g., AdvBench score)
- **Performance** → Median settling time and steady-state error vs reference KPI in production/shadow mode

These proxies are deliberately conservative and evidence-based. Institutions can customise but must document the mapping. The score is then auditable.

Determining and Applying Thresholds (The Real Decision Point)

Step 4 is where risk appetite meets operational reality. The threshold is not arbitrary; it is the explicit translation of business objectives, regulatory requirements, and risk tolerance for disruption into a single measurable number.

- A retail bank running a customer-service chatbot may accept an Overall Score ≥ 3.5 because the downside of failure is reputational, not existential.
- A Tier-1 investment bank running an algorithmic trading agent will demand ≥ 4.5 because a single instability event can cost hundreds of millions.

Choosing the threshold therefore requires a focused discussion with risk owners:

- What is the worst-case impact of loop failure (financial loss, regulatory breach, safety incident)?
- What is the frequency of disturbances we must withstand?
- What is the cost (latency, complexity, human overhead) of additional mitigations?

The chosen threshold is then documented in policy and reviewed at least annually or upon material change (new model version, scope expansion, regulatory update).

Proportionality ensures controls scale with risk: low-impact agents (e.g., internal chatbots) may operate at lower thresholds, while high-stakes ones (e.g., trading) demand stricter scores. This mirrors EU AI Act tiers, allowing innovation in sandboxes without over-governance.

Crucially, SAFE is designed for continuous monitoring, not one-off assessment. Scores are recomputed automatically on every model update, drift detection event, or quarterly review. When a score drops below threshold, the system automatically reduces autonomy (e.g., switches to human approval mode) until mitigations restore it.

The goal is to turn risk management from a paperwork exercise into a live engineering discipline.

How to quantify mitigation impact on scores

The uplift is deliberately semi-quantitative (engineering judgment + evidence), not purely mathematical, because real systems are too complex for perfect formulas. Typical process:

1. Baseline score from evidence (e.g., red-team success rate, measured latency, historical incidents).
2. Select mitigation(s) from proven catalogue (hierarchical supervision → controllability +1.0 typical).
3. Re-test or simulate with mitigation in place → measure new evidence → assign new score.
4. If needed, iterate.

Example: adding full CoT logging might raise Observability from 2 → 4 because red-team can now reconstruct 98 % of decisions (vs 30 % before). Adding a kill-switch might raise Controllability from 1 → 4 because override latency drops from minutes to milliseconds.

The key is transparency: every score change must be justified with evidence in the assurance statement.

SAFE Scorecard Template

Property	Score (1–5)	Evidence / Metrics	Mitigation Required?
Observability		% decisions with full CoT logging & provenance, reconstruction time (s)	
Controllability		Kill-switch latency (ms), escalation paths count	
Stability		Worst overshoot in last 20 scenarios (%)	
Robustness		Max tolerated adversarial prompt length	
Performance		Settling time (s), steady-state error (%)	

Overall Score formula (example – institutions customise weighting) = min(Obs, Cont) × 0.4 + (Stab + Rob + Perf)/3 × 0.6

Autonomy Level policy (customisable)[2]

Score	Autonomy Level	Typical Requirement
<2.0	0	Do not deploy / redesign
2.0–2.7	1–2	Human executes every action
2.8–3.4	3	Human approves high-impact actions
3.5–4.2	4	Human monitors + alerts
≥4.3	5	Full autonomy with minimal oversight

Walked-Through Examples

Example 1 – Loan Approval Agent

Step 1 – Mapping Reference: Approve good loans quickly while keeping portfolio PD < 0.8 % and NPS > 85 Controller: LLM credit analyst agent (up to SGD 500k authority) Environment: Credit portfolio + customer experience Sensors: Approval logs, repayment data, customer feedback

Step 2 – Raw assessment

Property	Raw Score	Failure mode without mitigation	Mitigation Applied
Observability	3	CoT logging present but incomplete for edge cases	Full CoT + provenance logging
Controllability	3	Human confirmation only > SGD 100k, kill-switch exists	Cryptographic commitment signatures on approvals > SGD 50k
Stability	3	Unclear behaviour under macro shocks	Daily automated replay of six historical macro shocks
Robustness	2.5	Vulnerable to adversarial borrower narratives	Adversarial red-team prompts every sprint
Performance	4.5	Very fast (favourable)	–

Raw overall ~2.9 → threshold set at 3.8 → mitigations applied → final score 3.8 → Level 4 granted (now live).

Example 2 – Multi-Agent Procurement System

Raw assessment (only Stability critically affected)

[2] Tailor levels to risk profile per ISO 42001 Annex A.6 (impact assessment).

Property	Raw Score	Failure mode	Mitigation Applied
Stability	1	Observed bidding wars (loop gain >> 1)	Supervisory budget agent + rate-limiting

Stability jumps to 4.5 → overall score 4.3 → Level 5 safe.

Example 3 – Human Trader

Same table format produces identical mitigations banks already use (pre-trade limits, four-eyes, mandatory leave). No AI exception needed.

Example 4 – Multi-Agent Fraud Detection

Raw assessment

Property	Raw Score	Failure mode	Mitigation Applied
Stability	2.5	Escalation spirals on ambiguous cases	Supervisory circuit-breaker on disagreement rate > 5 % in 10 min

Stability → 4.8 → overall score 4.1 → Level 4, −68 % fraud loss.

Example 5 – Agentic Browser / Personal Assistant ("Money Mules as a Service" risk)

Raw assessment

Property	Raw Score	Failure mode without mitigation	Mitigation Applied (raises to ≥4.2)
Observability	2	Hidden injections invisible; CoT may not reveal malice	Full CoT + provenance logging, real-time injection scanner
Controllability	1	Universal access across contexts	Hard context isolation, mandatory MFA for money movement, kill-switch
Stability	2	Single post causes catastrophic deviation	Rate-limiting + intent-change anomaly detector
Robustness	1	Known 2025 CVE-class indirect prompt injection	Input sanitisation, adversarial red-teaming
Performance	4	Very fast (favourable)	–

Raw overall ~1.8 → Level 0 With mitigation package → overall 4.4 → Level 5 safe

Same assessment applies verbatim to a human PA given identical access.

5. Relationship to Global Governance Frameworks (2025)

SAFE Property	ISO/IEC 42001:2023 (Annex A)	MAS FEAT / Nov 2025 Guidelines	NIST AI RMF / CSF 2.0	South Korea, Japan, China frameworks
Observability	A.8 Transparency, A.6.3 Monitoring	Transparency, Monitoring & Logging	Map, Measure	Transparency, Explainability, Watermarking
Controllability	A.9 Use, A.5.3 Human oversight	Accountability, Human Oversight	Govern, Protect	Human Oversight, Kill Switches, Emergency Response
Stability	A.6 Life cycle, A.5 Impact assessment	Fairness (systemic), Testing	Respond, Recover	Reliability, Robustness, Stability Testing

Robustness	A.6.4 Resilience	Ethics, Adversarial Robustness	Protect	Robustness, Security, Adversarial Training
Performance	A.6.3 Monitoring & measurement	Capabilities & Capacity	Measure	Performance Evaluation, Monitoring

Conclusion from table: ISO/IEC 42001 is the strongest certifiable standard. SAFE operationalises its required AI risk assessment and treatment clauses for agentic/dynamic risks, making certification substantially easier.

6. Multi-Agent and Hierarchical Extensions

Real-world 2025 deployments are rarely single-agent. Enterprise platforms already orchestrate dozens to thousands of concurrent agents (research agents, trading agents, customer-service agents, compliance-monitoring agents) that interact indirectly via shared markets, databases, calendars, or communication channels. These interactions routinely produce emergent phenomena: bidding wars, herding on stale signals, circular citation loops, apology cascades, and flash crashes — all classic symptoms of undamped positive-feedback across a fleet.

Control engineering solved large-scale interacting autonomous controllers decades ago in every safety-critical domain that actually ships (air-traffic control, power-grid frequency regulation, robotic swarms, high-frequency trading market-making). The solutions are mature, standardised, and directly applicable:

Control-theoretic concept	Translation to agent fleets (2025 practice)	Existing precedent
Hierarchical supervisory control	Slow human + fast automated supervisor that can throttle, pause, or revert any subset of agents in <500 ms	Aviation TCAS, nuclear SCRAM, HFT circuit breakers
Decentralised dissipativity / passivity-based design	Each agent emits a bounded "energy" signal (e.g. bidding intensity, P&L volatility); supervisor enforces dissipativity certificates across the fleet	Power-grid droop control, drone formation flying
μ-synthesis & structured singular value	Explicit robustness margins against worst-case interaction graphs (scale-free, small-world, or adversarial topologies)	Telecommunications network stability certification
Containment zones / invariant sets	Pre-computed "safe envelopes" that bound how far a misbehaving agent can propagate damage before kill layers activate	Autonomous vehicle responsibility-sensitive safety (RSS)

In SAFE terms, multi-agent risk is not a new category; it is simply a structured disturbance entering the loop at the interconnection layer. The same five properties are assessed at fleet level:

- Fleet observability → standardised telemetry schema (CoT fragments, tool-call rates, resource consumption) ingested by a supervisory controller in real time
- Fleet controllability → number and independence829 of kill layers, worst-case override latency
- Fleet stability → gain/phase margins under ensemble red-teaming (BIS-style herding scenarios, apology-loop induction, and persistence tests for self-modifying agents)
- Fleet robustness → maximum tolerated concentration before phase transition to instability (measured on Barabási–Albert graphs)

Organisations that skip these extensions are effectively running thousands of high-gain controllers in open-loop interaction — a configuration that no regulator would license in any other safety-critical domain.

7. Reference Integrity and Intent Alignment

The deepest alignment challenge is not "can the agent follow instructions?" but "does the closed loop remain tracking the right reference over indefinite horizons despite specification drift, proxy gaming, and ontological shifts?"

Control theory treats this as three well-understood sub-problems, each with eighty years of solutions:

Problem	Control-theoretic name	Required SAFE mechanisms (2025)
Proxy gaming / Goodhart	Sensitivity to reference misspecification	Distributionally robust reference design; explicit uncertainty sets for reward hacking
Long-term intent drift	Lack of integral action	Slow outer human loop that periodically re-asserts outcome-based reference (constitutional AI patterns, periodic board-level outcome audits)
Ontological shift (world model diverges from human values)	Environment uncertainty & time-variation	Adaptive control equivalents: model-reference adaptive systems, gain scheduling on world-model fidelity metrics, continual supervised fine-tuning on value-aligned trajectories

In practice this means:

1. Every reference must be explicit, version-controlled, human-auditable, and hierarchically decomposed (board → senior management → local agent) exactly as MAS November 2025 Guidelines and the EU AI Act now begin to require for high-risk systems.
2. Steady-state error must be provably bounded (integral-action equivalents): deviation from intended outcomes must trigger mandatory escalation, never silent accumulation.
3. Robustness margins must explicitly include specification-gaming adversaries (AdvBench-style reward-hacking prompts, DeceptionBench, and model-level specification gaming datasets count as structured uncertainty classes).

The reason LLM agents appear uniquely vulnerable to misspecification is that most current designs are pure proportional control with enormous gain and no outer integral loop — the textbook recipe for violent instability and Goodhart collapse[3]. Add the standard control-engineering fixes (hierarchical decomposition, slow integral outer loop, robust reference design) and the problem becomes tractable at scale.

8. Conclusion

Agent-Agnostic Risk Assessment Framework delivers the rigorous, measurable approach that regulators, boards and engineering leaders need to focus on as agentic AI becomes core infrastructure. Because it speaks the universal language of control engineering, it should remain valid long after today's LLMs. Most importantly, the framework completes the maturation of AI as a technology. Electricity, aviation, nuclear power and chemical processing all became ubiquitous only after control theory gave rigorous safety guarantees.

Agentic AI stands at the same threshold today — not just in finance, but in every domain where decisions have consequences. SAFE is deliberately a pragmatic engineering tool for systems we are deploying in 2025–2030. It complements, rather than replaces, longer-term alignment research aimed at potential superintelligent agents.

Consider healthcare: an agentic system that monitors patients, orders tests, adjusts medication doses, and coordinates care teams could transform outcomes and reduce costs dramatically. Yet without guaranteed loop stability, a single disturbance (sensor error, drug interaction discovery, or adversarial attack) could cascade into harm. SAFE forces the same disciplined conversation hospitals already have about pacemakers or infusion pumps: what score do we require before we grant autonomy? The answer determines whether the system stays in

[3] A Goodhart collapse is the dramatic, usually sudden failure mode that occurs when an optimization process pushes so hard on a proxy metric that the true underlying objective is catastrophically violated.

supervised mode (Level 3) or moves to full autonomy (Level 5) — exactly the conversation that turns experimental AI into trusted medical infrastructure.

Institutions adopting SAFE will find themselves naturally compliant with MAS, BIS, NIST, ISO 42001 and every other major regime, and will be the first to confidently deploy the full power of agentic systems at scale, because they can finally treat AI as just another extraordinarily useful, reliably safe technology.

The future is not only about better agents. It is closed-loop systems whose stability, observability and controllability we can measure and guarantee, no matter what kind of intelligence is inside the controller.

To evolve SAFE collaboratively, I welcome practitioner input on proxy refinements or domain adaptations—contact via www.davidroihardoon.com.

9. References

- Anthropic (2024). Building effective agents. https://www.anthropic.com/engineering/building-effective-agents (published 19 December 2024).
- Financial Stability Board (2024). The Financial Stability Implications of Artificial Intelligence. 14 November 2024. https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/
- Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., & Stein, M. (2024). Intelligent financial system: how AI is transforming finance. BIS Working Papers No 1194. Bank for International Settlements. https://www.bis.org/publ/work1194.htm
- FinRegLab (2025). The Next Wave Arrives: Agentic AI in Financial Services. September 2025. https://finreglab.org/research/the-next-wave-arrives-agentic-ai-in-financial-services/
- International Organization for Standardization (2023). ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. https://www.iso.org/standard/81230.html
- Monetary Authority of Singapore (2019). Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector. https://www.mas.gov.sg/publications/monographs-or-information-paper/2019/feat
- Monetary Authority of Singapore (2025). Consultation Paper on Guidelines on Artificial Intelligence Risk Management. P017-2025, November 2025.
- National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework
- National Institute of Standards and Technology (2024). The NIST Cybersecurity Framework (CSF) 2.0. NIST CSWP 29. https://doi.org/10.6028/NIST.CSWP.29
- People's Republic of China National Information Security Standardization Technical Committee (TC260) (2025). Artificial Intelligence Safety Governance Framework 2.0. September 2025.
- Republic of Korea (2025). Framework Act on Artificial Intelligence (promulgated January 2025, effective 2026).
- Ministry of Economy, Trade and Industry, Japan (2025). AI Guidelines for Business Version 1.1. April 2025. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_14.pdf
- Additional foundational references: OECD AI Principles (2019), UNESCO Recommendation on the Ethics of Artificial Intelligence (2021), EU AI Act (2024).
- Guo et al. (2024). ControlAgent: Automating Control System Design via Novel Integration of LLM Agents and Domain Expertise. arXiv:2410.19811
- Zhang et al. (2025). A Control-Theoretic Approach to Generative AI Guardrails. arXiv:2510.13727
- Jia et al. (2025). AgenticControl: An Automated Control Design Framework Using LLM Agents. arXiv:2506.19160
- Zahedifar et al. (2025). LLM-Agent-Controller: A Universal Multi-Agent Large Language Model System as a Control Engineer. arXiv:2505.19567