From AI Pilot to Sustainable Value: A Prudential Framework for AI Commercialisation

Amit Godbole

Independent Researcher, India

DOI: https://doi.org/10.51584/IJRIAS.2026.11090112

Received: 24 September 2026; Accepted: 29 September 2026; Published: 19 October 2026

ABSTRACT

Surveys show organisations adopting artificial intelligence (AI) faster than they report earnings from it: the share using AI rose from 78 per cent in 2024 to nearly nine in ten in 2026, while the share attributing any earnings impact stayed near 37 per cent. Research on business models, scaling and governance does not connect its evidence to the decision to commit more to an initiative. This paper develops the AI Commercialisation Prudential Framework (AICAF) to make that connection. Its principle is that commercial exposure should rise only when the evidence on six adequacy dimensions meets a floor set by the gate and an exposure rating. The design borrows four mechanisms from Basel II and from supervisory model-risk guidance (SR 11-7, issued in 2011 and since superseded, and PRA SS1/23): minimum conditions set in advance, independent challenge, a recorded assumption set, and an evidence burden that rises with what is at stake. Capital has no counterpart, and without a mandate to enforce the gate the framework is advisory. A purposive structured review of five literatures yields the six dimensions; a sixth, on decision under uncertainty, supplies the rationale for a hold. The framework has four evidence levels with a non-compensatory rule, a three-tier exposure rating, five gates and four decision outcomes: advance, hold, remediate or stop. A calibration rubric and a lean version support its use. I applied its dimensions and outcomes retrospectively to eight initiatives that I led or advised, excluding two whose later outcome could not be established; every reading is my own, and exposure tiers and evidence levels were not recorded. A retrospective reading of Zillow Offers and GitHub Copilot illustrates the gates. The framework is compared with Stage-Gate, SPARK, Lean Startup, MLTRL and the NIST AI RMF. Six propositions and five validation designs remain as future work.

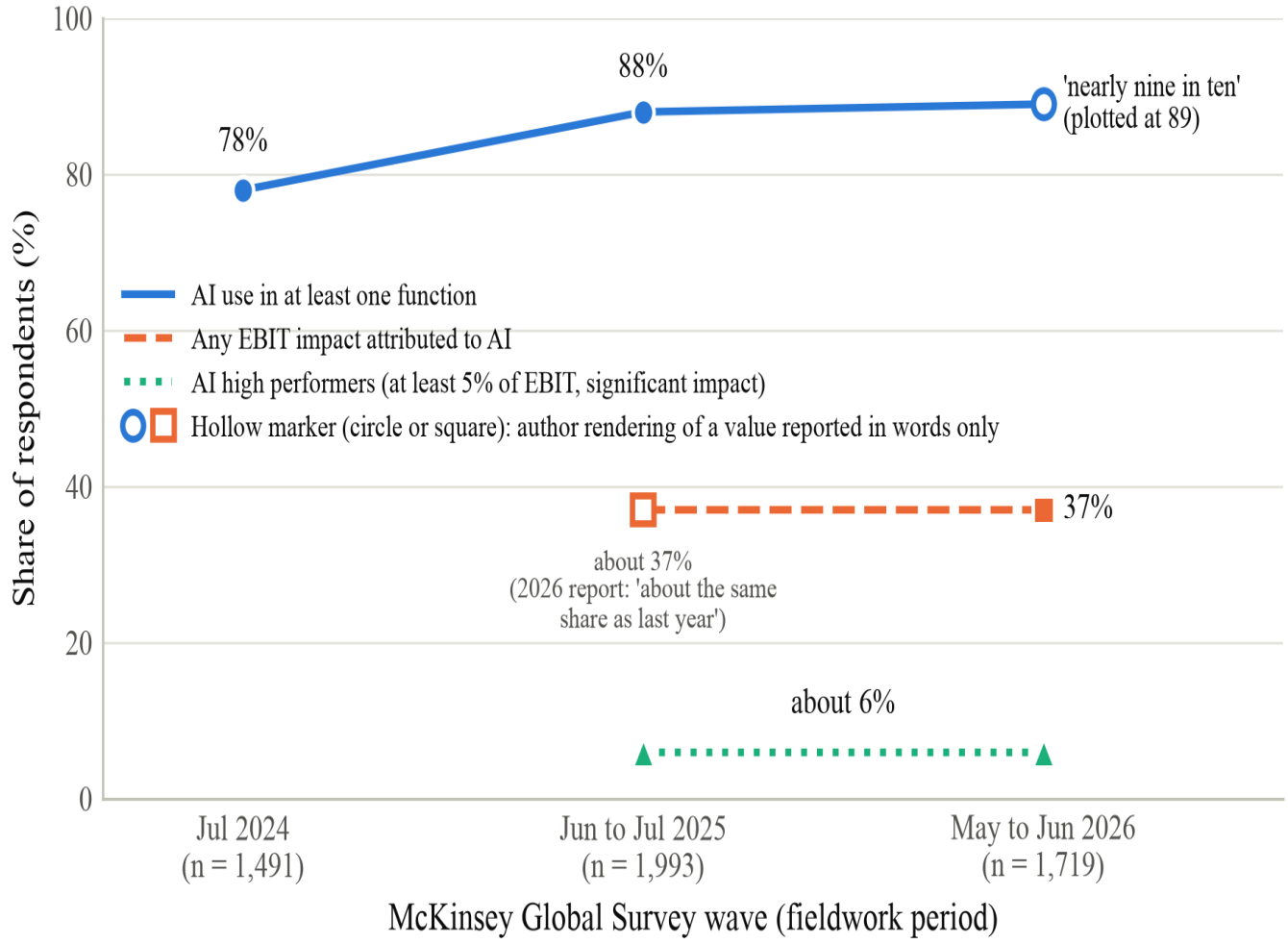
Keywords: AI commercialisation, evidence adequacy, exposure maturity, business model innovation, AI governance

INTRODUCTION

In McKinsey's Global Survey waves fielded in July 2024, June to July 2025 and May to June 2026, the share of organisations using AI in at least one business function rose from 78 per cent to 88 per cent to nearly nine in ten [63,64,65]. In the 2026 wave 37 per cent of respondents attributed at least some impact on earnings before interest and taxes (EBIT) to AI, about the same share as the year before, and the share of "AI high performers", firms attributing at least 5 per cent of EBIT to AI and describing the impact as significant, stayed at about 6 per cent [65]. Figure 1 places the three series side by side. Investment moved on a different scale: global corporate AI investment reached USD 581.69 billion in 2025, 129.9 per cent above the 2024 figure [85]. Figure 2 sets that series against the EBIT share as an annotation; the two describe different populations, so the figure shows a divergence and no cause.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

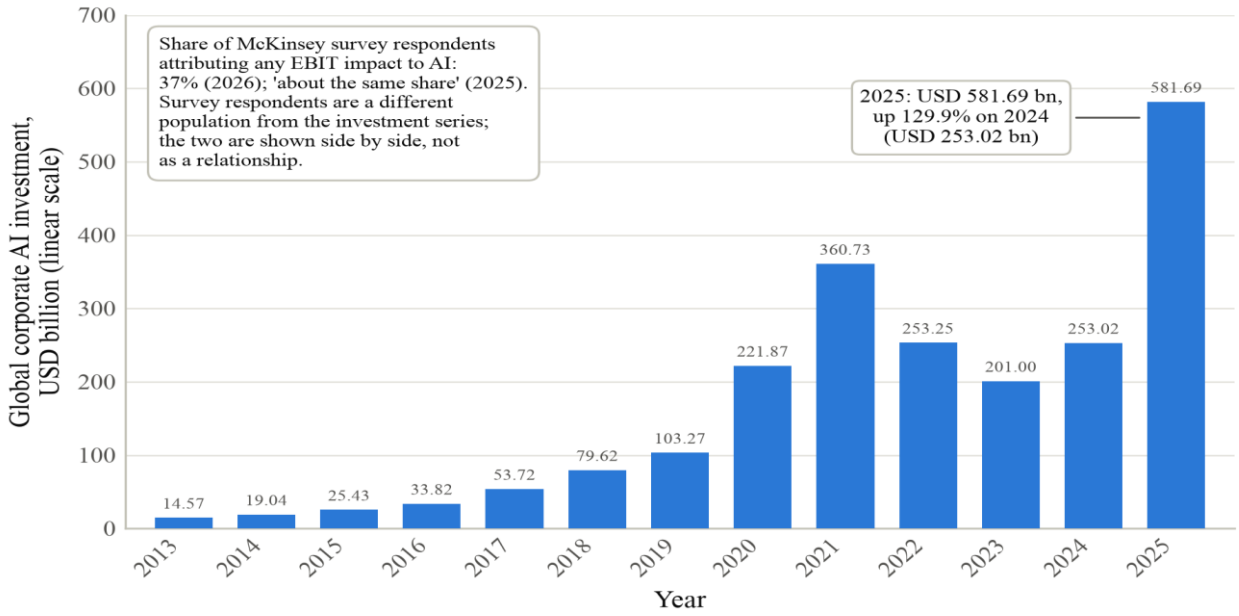
Figure 1: Organisational AI adoption against reported earnings impact, McKinsey Global Survey waves fielded July 2024 (n = 1,491), June to July 2025 (n = 1,993) and May to June 2026 (n = 1,719). Adoption is the share of respondents whose organisations use AI in at least one business function; EBIT impact is the share attributing at least some EBIT impact to AI; high performers attribute at least 5 per cent of EBIT to AI and describe the impact as significant. Hollow markers are author renderings of verbal reports (Table 15, Appendix B). No 2024 values are published for the EBIT and high-performer series, and earlier waves are omitted because the high-performer definition changed.



Note: no 2024 values are published for the EBIT and high-performer series. Earlier waves are omitted because the high-performer definition changed.

Source: McKinsey & Company, 2025 and 2026 [63,64,65].

Figure 2: Global corporate AI investment, 2013 to 2025 (USD billion, linear scale), with the share of organisations attributing any EBIT impact to AI shown as a text annotation. Investment aggregates mergers and acquisitions, minority stakes, private investment and public offerings as compiled by Quid; the AI Index chapter does not state whether the series is in nominal or constant dollars. The EBIT share comes from a survey of enterprise respondents, a different population, and is not plotted as a series.



Source: Stanford AI Index Report 2026, Chapter 4, Figure 4.2.1 [85]; McKinsey & Company, 2026 [65].

Surveys at the project level report abandonment too, in two measures that must be kept apart. S&P Global's 2025 survey of 1,006 professionals found the share of organisations abandoning most of their AI initiatives before production rising from 17 per cent in 2024 to 42 per cent in 2025, with 46 per cent of projects on average scrapped between proof of concept and broad adoption [84]. Gartner's measure is narrower: it predicted in July 2024 that at least 30 per cent of generative AI projects would be abandoned after proof of concept by the end of 2025 [34], and in January 2026 reported that at least 50 per cent had been, without a published sample size [16]. The stated causes are not model failures. RAND's interviews with 65 data scientists and engineers identified five root causes, beginning with misunderstanding of the problem to be solved [80], and Cooper, the originator of Stage-Gate, finds that most business AI projects fail for the same business-practice reasons as new products [19]. Even in a supervised sector, firms use AI they do not fully understand: in the Bank of England and Financial Conduct Authority's 2024 survey of 118 financial firms, 75 per cent used AI and 34 per cent reported a complete understanding of it [5].

I read these figures as evidence of a decision problem; that reading motivates the paper and is not a finding of it. Each step from experiment to pilot to production to scale raises the organisation's exposure to cost, operational dependency, customer expectation, regulatory consequence and reputational damage, and the evidence needed to justify the next step should rise with it. Six streams of literature each hold part of the answer: technology readiness, with the scaling and deployment studies that the Literature Review treats under a heading of their own [53,57,68,75,91], business-model research [3,86], staged investment and validated learning [1,60,81], decision theory [42,89], AI governance [44,69], and prudential and model-risk supervision [29,77]. None supplies a bridge from that evidence to the economic decision to increase or reduce exposure that a committee without agreed probabilities can use. Large reviews of AI-driven business model innovation, covering 180 articles [48] and 87 articles [46], and a 2026 review of 98 articles on AI in new product development [88], describe no decision rule that links stage, evidence and exposure.

The framework applies to two kinds of organisation. An enterprise deploying AI inside its own operations raises its exposure as more processes and people depend on the system; a firm selling an AI product raises its exposure as more customers, revenue and contractual obligations depend on it. The paper calls the first an adopter and the second a vendor, and "AI initiative" covers both: a bounded application or product investment that moves through identifiable stages of development and deployment. The EU AI Act's provider and deployer roles are a different distinction, which the Discussion keeps apart. "Sustainable value" in the title means value that survives the move from pilot to production and scale; the word carries no environmental meaning here.

Prudential supervision solved a structurally similar problem for banks: Basel II set minimum capital requirements, a supervisory review process and market discipline through disclosure [6], and later model-risk

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

guidance added independent validation and controls tiered by materiality and complexity [29,77]. I do not argue that AI products should be regulated like banks; the Literature Review states which parts of that architecture carry across and which, capital first among them, do not.

The research question is: how can an organisation determine whether an AI initiative has generated sufficient evidence to justify increasing its commercial exposure? I answer it by synthesising five of the streams named above into the AI Commercialisation Prudential Framework (AICAF) and reading the result back against the sixth, decision under uncertainty.

The paper makes four contributions. It brings technical, commercial, operational and governance evidence into one decision architecture. It defines evidence adequacy as an ordinal construct with four levels and a non-compensatory aggregation rule. It defines exposure independently of stage, as a rating that sets the evidence floor at each gate. And it distinguishes technology maturity from exposure maturity, which explains why a technically advanced AI initiative can remain commercially under-evidenced. Three supports accompany the contributions and are smaller claims: a calibration rubric with a rater-agreement procedure (Appendix C), whose thresholds are illustrative; a retrospective application of the framework's dimensions and outcomes to eight initiatives that I led or advised, with the limits the Findings state; and a retrospective reading of two public cases, Zillow Offers and GitHub Copilot, as secondary analysis with the outcomes known. Controlled testing of the propositions remains future work.

LITERATURE REVIEW

Technology readiness measures capability, not commercial readiness

Technology readiness scales are the natural starting point because they separate scientific possibility from demonstrated maturity. NASA's Technology Readiness Level (TRL) scale, first set out in Mankins's 1995 white paper and now codified in Appendix E of NPR 7123.1D, runs from TRL 1, basic principles observed, to TRL 9, a system proven in operation [57,68]. Lavin and colleagues adapted the idea to machine learning as MLTRL, a ten-level scale from level 0, first principles, to level 9, deployment [53].

Two peer-reviewed critiques limit how far a readiness level can be trusted as a decision input. Olechowski and colleagues identified 15 challenges to TRL assessment [74]. Martínez-Plumed, Gómez and Hernández-Orallo reinterpreted the nine levels for AI and added a second axis, generality: high readiness is reachable for narrow capabilities while remaining out of reach for general ones [59]. A readiness level answers one question, whether the technology is mature enough to function in its intended environment, and says nothing about the customer, the workflow or the risk the organisation will carry. Exposure maturity, the construct proposed here, rates the organisation's evidence against what it is about to put at risk.

Commercial readiness needs those other answers. Teece treats the business model as the firm's design for creating value for customers, delivering it and converting part of it into profit, which makes technical feasibility and commercial viability separate questions [86]. Sjödin and colleagues showed that scaling AI in manufacturing firms required AI capabilities and business-model changes to co-evolve through feedback loops [82], and Güner Gültekin and colleagues' interview study of ten AI entrepreneurs in Turkish technoparks found no deliberate approach or model for commercialisation among them [38].

Business model innovation and value capture

Zott, Amit and Massa established the business model as a distinct unit of analysis linking value creation to value capture [102], and Amit and Zott showed that new value can come from redesigning the activity system without any change in technology [2]. AI sharpens the value-capture side: Åström, Reim and Parida give value-capture mechanisms a phase of their own in their process for AI business model innovation [3], and an AI offering's cost of delivery is coupled to inference, data acquisition, human review and external model suppliers. Teng, Ye and Martinez, from 32 interviews in the information technology industry, found that foundation models enable several kinds of business model innovation but rely on upstream support, so the downstream firm's economics depend partly on a supplier's pricing and terms [87]. Kanbach and colleagues' scoping review expects generative

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

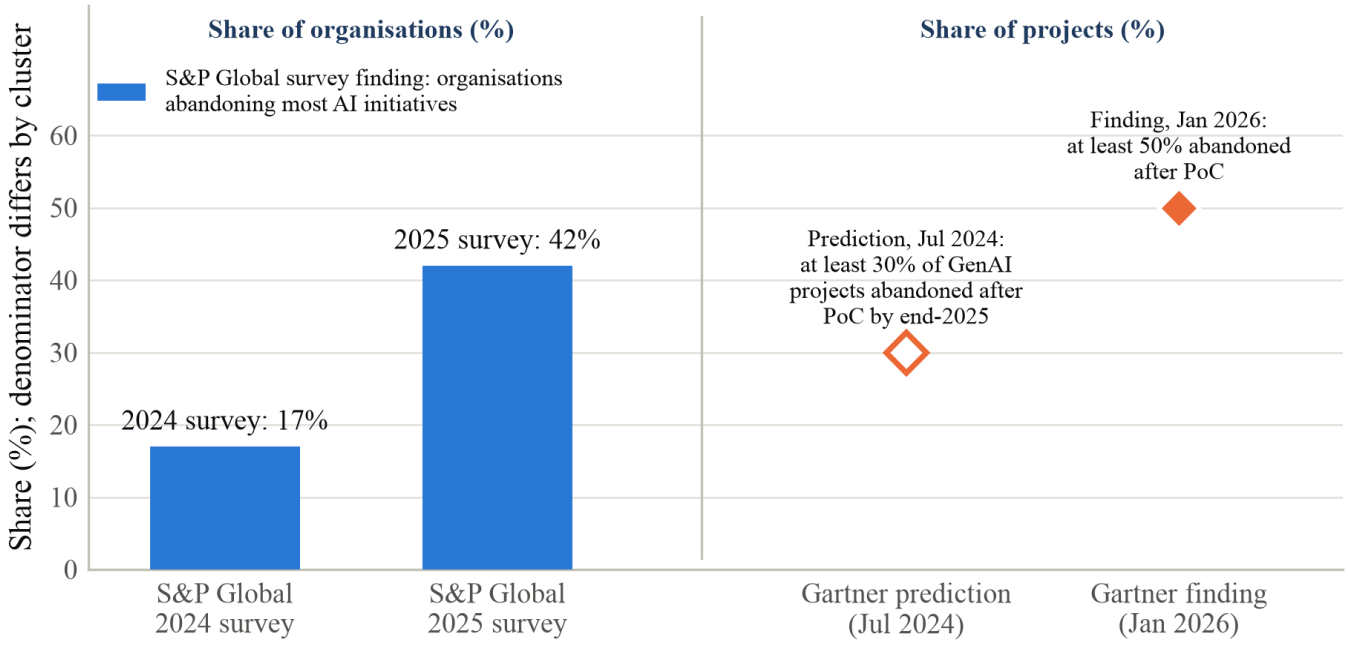
AI to reshape value propositions and cost structures across industries [50]. A pilot can thus demonstrate value creation while leaving value capture untested, and AICAF gives value capture its own test, economic adequacy.

Scaling, deployment and the pilot-to-production problem

Scaling adds a third layer of uncertainty. MIT CISR, investigating 52 AI solutions since 2019, distinguishes "scaling up", in which a solution moves from model development to pilot to production with increasing value creation, from "scaling out", in which a model is trained on new data and new expertise for a new related application [91]. Paleyes, Urma and Lawrence, surveying published deployment case studies, found practitioners facing distinct problems at every stage of the deployment workflow [75]. Haakman and colleagues, in 17 interviews at the bank ING, found that standard AI lifecycle models such as CRISP-DM omitted stages the bank had to add in practice, among them model monitoring and model risk assessment [39]. Risk assessment and monitoring therefore belong inside the lifecycle as stages with their own evidence, as the NIST framework also holds [69].

Figure 3 shows the two abandonment measures from the Introduction, each with its own denominator, and Figure 4 shows where pilots stall, with each measure labelled by its own definition. The MIT NANDA report in Panel A is often summarised as "95 per cent of AI pilots fail"; the 95 per cent is the share of organisations whose generative AI pilots showed no measurable profit-and-loss impact. Its pilot funnel is a separate exhibit: 60 per cent of organisations investigated embedded or task-specific generative AI tools, 20 per cent piloted them and 5 per cent reached production [15]. The 46 per cent of proofs of concept reaching production in Panel B and the 46 per cent of projects scrapped in Figure 3 are the same number with opposite meanings from different samples.

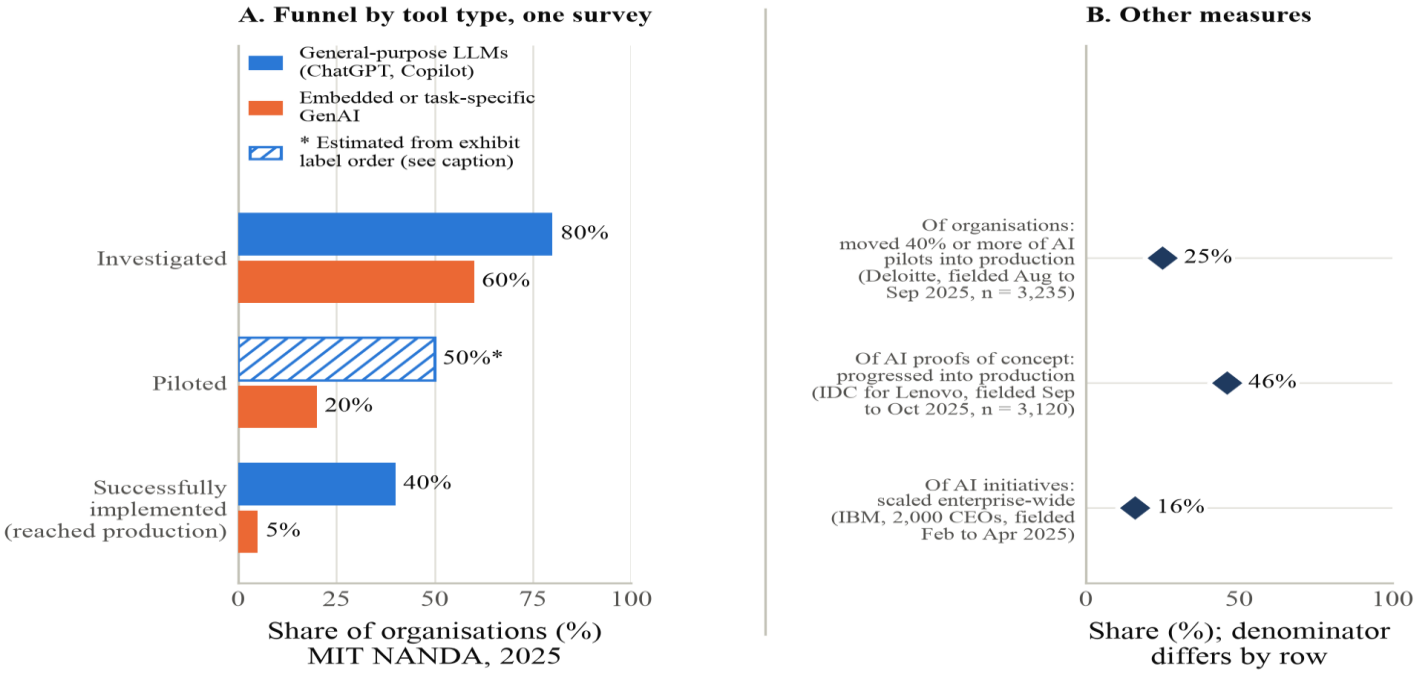
Figure 3: Two abandonment measures, each with its own definition and denominator. Bars, share of organisations that abandoned the majority of their AI initiatives before production: 17 per cent in the 2024 survey and 42 per cent in the 2025 survey (S&P Global Voice of the Enterprise, n = 1,006). Markers, share of generative AI projects abandoned after proof of concept: Gartner's July 2024 prediction of at least 30 per cent by end-2025 (hollow) and its January 2026 finding of at least 50 per cent (filled), neither with a published sample. The two constructs are not one series.



Bars: share of organisations abandoning the majority of their AI initiatives before production. Diamonds: share of GenAI projects abandoned after proof of concept (hollow = prediction, filled = later finding). The two constructs differ and are not one series. The S&P Global 2025 survey also reported an average 46% of projects scrapped between proof of concept and broad adoption.

Source: S&P Global Market Intelligence, 2025 [84]; Gartner, 2024 [34]; Chandrasekaran, 2026 [16].

Figure 4: Where AI pilots stall, by source and definition. Panel A: share of organisations that investigated, piloted and reached production with general-purpose and task-specific generative AI tools (MIT NANDA, July 2025; read from a publicly mirrored copy of the v0.1 report, as reference [15] states). The general-purpose 'Piloted' value of 50 per cent, asterisked and hatched, is the author's estimate read from the ordering of labels on the report's exhibit. Values are shares of organisations, not of pilots. Panel B: three measures with different definitions and denominators, not comparable with Panel A or with each other: 25 per cent of organisations moved 40 per cent or more of AI pilots into production (Deloitte); 46 per cent of AI proofs of concept progressed into production (IDC for Lenovo); 16 per cent of AI initiatives scaled enterprise-wide (IBM). Samples and fieldwork periods are in Table 15, Appendix B.



Source: Challapally et al., 2025 [15]; Deloitte, 2026 [24]; Lenovo, 2026 [55]; IBM Institute for Business Value, 2025 [43].

Product-development research offers the oldest answer to staged commitment. Cooper's Stage-Gate positions gates as decision points with defined gatekeepers and explicit success criteria, and his 2008 update describes scalable versions, Stage-Gate XPress and Stage-Gate Lite, in which the number of gates and the deliverables are set by project risk and size [18]. Gate criteria are of two kinds: must-meet criteria, which are knock-out questions answered yes or no and are non-compensatory, and should-meet criteria, which are scored and weighed against one another [18]. AICAF's evidence floor is modelled on the must-meet criterion, and the Discussion states what it adds. Cooper and Sommer's Agile-Stage-Gate hybrid runs Agile sprints inside stages while keeping gates as decision points [21]. Cooper's 2025 SPARK model proposes a scoring model inside a gated process for go/kill decisions on AI projects, arguing that scoring counteracts intuition bias [20]; AICAF shares its diagnosis and differs on aggregation, for reasons given in the Findings.

Staged investment, validated learning and escalation

A fourth stream supplies the theory behind the gates. McGrath framed early technology investments as real options: small, staged commitments buy the right, without the obligation, to scale once uncertainty resolves, and the value of that right rises with uncertainty [60]. Her later work treats the failure of individual initiatives as an expected and informative outcome whose cost should be contained while its learning is preserved [61]. Adner and Levinthal set the boundary condition. Real-options discipline only works where abandonment is enforced; when a firm can redefine an initiative's goals mid-way, the option logic collapses and escalation of commitment follows [1]. Sleesman and colleagues' meta-analysis of escalation determinants shows how common that collapse is [83].

The entrepreneurship literature holds the nearest thing to causal evidence that evidence-gated decisions improve outcomes. The lean startup was set out for practitioners by Ries [79] and Blank [12], and Shepherd and Gruber connect its "validated learning" to the literature on hypothesis testing and pivoting [81]. Felin and colleagues state its boundary: the method orients entrepreneurs towards ideas that can be tested quickly and transparently with customers, which prompts a search for validation where it is easy to observe and favours incremental experiments [31]. Camuffo and colleagues' randomised trial with 116 Italian start-ups found that founders trained to treat their ideas as hypotheses and test them rigorously performed better and were more willing to pivot [13], and a replication across four trials and 759 firms found that the approach raised idea termination [14]. Evidence adequacy is the commercial counterpart of validated learning: the point at which a hypothesis about value, operations or economics has been tested well enough to justify the next commitment.

Decision under uncertainty: sequential tests, information value and adaptive plans

The gate has an ancestry older than product development. Wald's sequential analysis defined a test whose rule returns one of three decisions at every stage of an experiment: accept the hypothesis, reject it, or continue by taking another observation [89] (p. 118), so the amount of evidence is not fixed in advance. Advance, Stop and Hold correspond to those three options, loosely: Wald's two terminal decisions end the test, whereas Advance opens the next gate, and his rule rests on one statistic where a gate joins several dimensions. Remediate has no counterpart in Wald.

Decision theory explains why the third option can be the economically right one. Howard showed that a numerical value can be assigned to reducing an uncertainty, that the value depends on the consequences of the decision as well as on the probabilities, and that the value of perfect information is an upper bound on what any experiment aimed at the same uncertainty can be worth [42]. On that logic a Hold is a purchase of information, justified when the evidence awaited could change the decision and is worth more than it costs, the cost of delay included. The gate rule does not weigh that cost; the false-Hold discussion is where it enters the framework.

Magee brought the reasoning to capital investment through the decision tree, which sets out the choices, risks and information needs of a staged decision [56]. Dixit and Pindyck built a theory of investment on three features most investment decisions share: the outlay is at least partly irreversible, the rewards are uncertain, and the investor has some leeway over timing. Where all three hold, the ability to wait for information has a value that is given up on investing [25]. The real-options work reviewed above concerns the option to grow, bought with a small early commitment [60,61]; Dixit and Pindyck's is the option to defer, and it is one of the reasons cited here for a floor that rises with exposure. It applies to the people and money components of Table 4, inside which reversibility is rated; a tier set by autonomy or regulatory class rests on the consequence-based reasoning of Howard and of model-risk tiering [42,77].

Knight separated measurable uncertainty, which he called risk, from the unmeasurable kind [51], and in my judgement the commercial outcomes of an early AI initiative belong to the second. The literature on deep uncertainty plans for that condition [54,58]. In Haasnoot and colleagues' dynamic adaptive policy pathways, signposts specify the information to be tracked, triggers are the critical values of a signpost beyond which further action is taken, and an adaptation tipping point is the point at which an action no longer meets the plan's objectives [40]. In AICAF the monitored measures are the signposts and the thresholds set at G3 are the triggers. The change events that reopen a gate after production are closer to adaptation tipping points, since a change is material when the evidence on record no longer describes the system as it will run.

AICAF takes from this literature the option to continue gathering evidence, the principle that the worth of evidence depends on what is at stake, and the practice of committing in short steps with triggers set in advance. It attempts no quantitative optimal-stopping model: Wald's thresholds come from stated error probabilities, and information-value and option-value calculations need probabilities and payoffs that a committee facing an early AI initiative usually lacks. The framework is an ordinal decision rule, related to the formal literature by lineage and analogy and not derived from it. The literature justifies the existence of Hold; where Table 6 places the floors remains a convention.

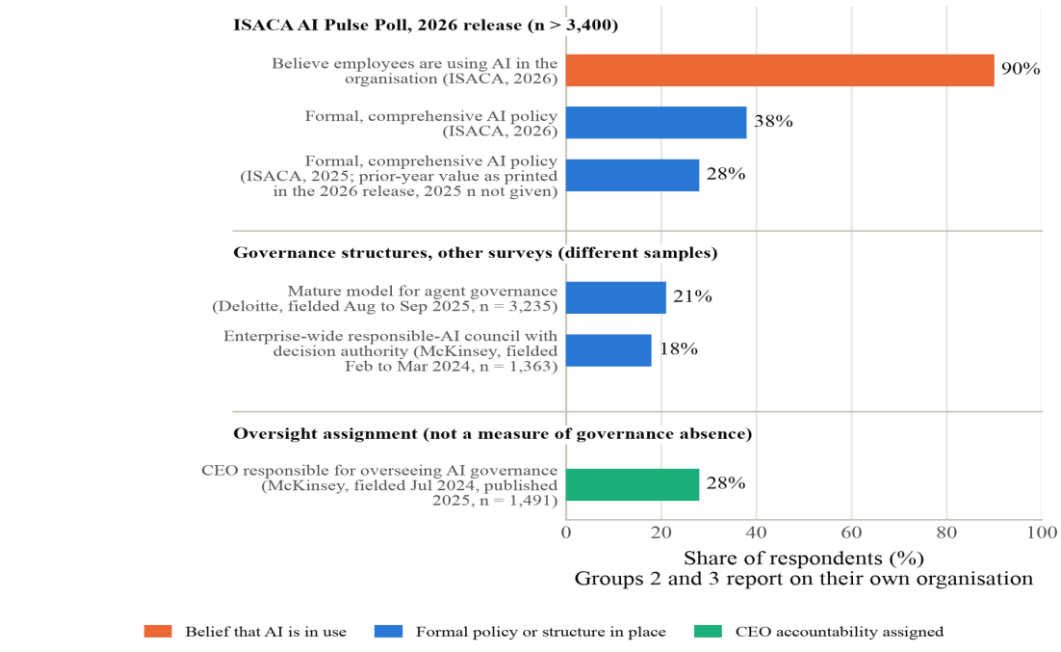
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

AI governance and risk frameworks

Governance forms a further stream. The NIST AI Risk Management Framework organises AI risk management into four functions, Govern, Map, Measure and Manage, across the system lifecycle [69,70], and its Manage function comes closest to a commercial decision. Subcategory MANAGE 1.1 of AI RMF 1.0 [69] reads: "A determination is made as to whether the AI system achieves its intended purposes and stated objectives and whether its development or deployment should proceed" (the Playbook entry [71] prints "purpose" in the singular). The framework calls for the determination but supplies no economic criterion for making it. NIST's 2024 Generative AI Profile lists twelve risks unique to or exacerbated by generative AI [72]. ISO/IEC 42001:2023 sets requirements for an AI management system covering AI-related risks and opportunities across the lifecycle [44]. The European Union's AI Act adds a legal tier structure: prohibited practices, high-risk systems, transparency obligations and minimal-risk systems, plus a separate regime for general-purpose AI models [27]; it allocates obligations by role, and Regulation (EU) 2026/1744 revised its application dates in July 2026 [28]. The Discussion develops both points.

On my reading these instruments were designed for responsible AI management; allocating commercial investment was never their purpose. They also lag use, as Figure 5 shows: in ISACA's 2026 poll of digital-trust professionals, 38 per cent of organisations had a formal, comprehensive AI policy, up from 28 per cent a year earlier, while 90 per cent of respondents believed employees were using AI [45]; other surveys put enterprise-wide responsible-AI councils with decision authority at 18 per cent [62] and mature governance of AI agents at 21 per cent of companies planning to deploy them [24]. Raisch and Krakowski explain part of the lag: automation and augmentation are interdependent over time, so the governance of a human-in-the-loop workflow cannot be settled once at deployment [78]. The gap this paper addresses sits between the governance evidence these frameworks produce and the economic decision to raise or lower exposure.

Figure 5: Formal AI governance against AI use, in three groups from different surveys. Group 1 (ISACA AI Pulse Poll, 2026 release): 90 per cent of respondents believe employees use AI in their organisation, a belief measure, against 38 per cent whose organisations have a formal, comprehensive AI policy, with 28 per cent as the prior-year comparison printed in the same release. Group 2 (governance structures, different samples): 21 per cent of companies planning to deploy agentic AI report a mature model for agent governance (Deloitte); 18 per cent report an enterprise-wide responsible-AI council with decision authority (McKinsey). Group 3: 28 per cent report that the CEO oversees AI governance (McKinsey), which records where accountability sits and is not evidence of a lag. Samples and fieldwork periods are in Table 15, Appendix B.



Source: ISACA, 2026 [45]; McKinsey & Company, 2024 and 2025 [62,63]; Deloitte, 2026 [24].

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Prudential logic, Basel and model risk

Basel II set out three pillars: minimum capital requirements, a supervisory review process, and market discipline through disclosure [6]. Basel III revised the capital and liquidity content without changing that architecture [7], and the standards are now a consolidated Basel Framework [8]. Under Pillar 2, chapter SRP20 sets four principles: the bank's own assessment of its capital adequacy against its risk profile (principle 1), the supervisor's review of that assessment (principle 2), the expectation that banks operate above the minimum, with the supervisor's "ability to require banks to hold capital in excess of the minimum" (principle 3), and early intervention before capital falls below the minimum (principle 4) [9]. The same chapter requires forward-looking stress testing [9], and the disclosure standard, DIS, sets out the requirements that aim to encourage market discipline [10].

Model-risk supervision offers a tighter analogue that does not depend on the capital metaphor, because it gates the use of a model on validation evidence. SR 11-7, the guidance issued in 2011 by the Federal Reserve and the Office of the Comptroller of the Currency, rested on three elements: sound model development, implementation and use; validation independent of the model's owners, comprising conceptual soundness, ongoing monitoring and outcomes analysis; and governance, policies and controls, each proportionate to the model's materiality and complexity [29]. The US banking agencies superseded it on 17 April 2026 with revised interagency guidance, SR 26-2, which keeps effective challenge, the same validation components and rigour commensurate with model risk, places generative and agentic AI models outside its scope, and states that it sets no enforceable standards [30]. I cite the 2011 text as the source of the design principle and make no claim about current US supervisory practice. The Prudential Regulation Authority's SS1/23 sets five principles, beginning with model identification and risk classification and including independent validation, and states that it covers the risks of using AI techniques in modelling [77]; its April 2026 update clarifies that its expectations are not conditions for internal model approval and changes nothing relied on here. Subject to that US exclusion of generative and agentic models, supervisors already route AI through this logic: the BIS Financial Stability Institute reports that most financial authorities have not issued AI-specific regulation because existing frameworks address most of the risks [22], and the Financial Stability Board names third-party dependency and supplier concentration among the main AI vulnerabilities [32].

The analogy is narrow, and Table 1 states its limits row by row. Capital, the mechanism at the centre of Basel, has no counterpart: nothing in AICAF absorbs a loss once exposure has been approved. Pillar 3 does not transfer either, because market discipline depends on outside parties pricing the disclosed risk with their own money at stake. What carries across is the approval architecture, four mechanisms in all: a floor set in advance and judged dimension by dimension, challenge by a party that does not own the initiative, a record of what would invalidate the case, and a burden that rises with what is at stake. A prudential approach to AI commercialisation, as the term is used here, is a decision discipline built on those four mechanisms.

A second limit concerns enforcement. The Basel standards take legal force through their implementation in national or Union law, and under that mandate the supervisor reviews the bank's own assessment, can require capital above the minimum and can intervene before capital falls below it, as SRP20 describes [9]. Most enterprises have no such enforcer, and without one a gate decision binds nobody. Four arrangements can substitute in part. Two are preventive and can block an Advance: a board or investment committee that makes funding conditional on the gate record, and covenants in customer contracts or investor agreements that require the evidence. Internal audit is detective: it tests afterwards whether gates were held and overrides recorded. Supervisory expectations do the work where they apply, as SS1/23 does for the banks in its scope and as SR 11-7 did until it was replaced; SS1/23 is, and SR 11-7 was, supervisory guidance, applied through examination and the supervisor's general powers [29,77], and the 2026 US guidance states that non-compliance with it "will not result in supervisory criticism" [30]. Even with a mandate the challenge is weaker than a supervisor's: the decision forum may Advance against the challenger's level with a recorded reason, a comply-or-explain arrangement, whereas a bank cannot override its supervisor. Where no mandate exists, where sponsor and challenger share a reporting line, or where a founder holds both roles, AICAF is advisory. It can still structure a decision, but P5 and P6 (Discussion), the propositions about challenge and a rising floor, should not be expected to hold there, and a test of them should exclude such settings or treat them as a comparison group. P1 does not depend on enforcement and finds its below-floor group in these settings.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Table 1: Prudential mechanisms and their AICAF counterparts, if any.

Prudential mechanism	Source	AICAF counterpart, if any	What is transferred, and what is not
Minimum capital requirements (Pillar 1)	Basel II [6]	None for the buffer. The design principle survives as the evidence floor per gate and exposure tier (Table 6)	Only the principle that a floor is set before exposure is taken. The floor's non-compensatory property comes from Stage-Gate's must-meet criterion, not from Pillar 1, under which capital is fungible across exposures.
Supervisory review and early intervention (Pillar 2, SRP20 principles 2 and 4)	Basel Framework [9]	Independent challenge by a party outside the sponsoring unit's reporting line, who records a level that may differ from the owner's; the owner's level corresponds to the bank's self-assessment under principle 1	Challenge recorded before exposure rises. The challenger records and the forum decides, and the forum may override with a recorded reason, which a bank cannot do to its supervisor. The challenger is internal or contracted, not a public authority.
Forward-looking stress testing (SRP20)	Basel Framework [9]	Assumption record and sensitivity questions attached to each decision	The obligation to state what would invalidate the case.
Disclosure and market discipline (Pillar 3)	Basel Framework [10]	None. The assumption record is visible to the decision forum	Not transferred. Internal visibility to leadership is documentation, not market discipline.
Effective challenge and independent validation, proportionate to materiality and complexity	SR 11-7 [29], issued in 2011 and superseded in 2026 [30]; SS1/23 [77]	Two independence criteria for the challenger; evidence floor set by exposure rating (P6, Discussion)	Independence and proportionality. AICAF departs from the sources by not using model complexity as a classifier, because a supplied foundation model of unknown complexity can carry high exposure and a complex in-house model can carry little.
Model risk classification by tier (materiality and complexity in the source)	SS1/23 [77]	Three-tier exposure rating (Table 4)	Tiering of the evidence burden by consequence.
Ongoing monitoring and outcomes analysis as validation components	SR 11-7 [29]	Monitoring cadence, change control and decommissioning after production (Findings, "After production")	The requirement that evidence be maintained, not only produced once.

METHODOLOGY

This is a conceptual paper built by a purposive structured review and framework development, with a retrospective application of the framework's dimensions and outcomes to eight initiatives, as the Findings report.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

The objective is a decision architecture, built from established concepts, that can later be tested; the paper does not estimate an effect and does not claim exhaustive coverage of the literature.

Source identification and inclusion

Sources were identified in three ways: searches of Crossref, Semantic Scholar and Google Scholar on the terms of the six streams; backward and forward citation chaining from seed papers; and direct retrieval of institutional documents from standard setters, legislators and supervisors. Each candidate was checked against the publisher's or standard-setter's copy before inclusion, with the exceptions listed in the source-verification note in the Declarations, and candidates that could not be confirmed were discarded. Table 2 states the criteria. The sixth stream, decision under uncertainty, entered the review after the journal's reviewer identified its absence; its sources are foundational works reached by citation chaining, and none was coded for conditions. Screening counts were not logged. The reference list holds 102 sources: 30 coded for conditions, 32 survey, index and company-document sources, and 40 context sources; the note under Table 3 lists the three groups by reference number.

Table 2: Inclusion and exclusion criteria for the review.

Criterion	Include	Exclude
Period	2010 to September 2026; earlier foundational works where later work still builds on them	Earlier works that later work no longer builds on
Coded sources: type	Peer-reviewed articles and primary institutional documents (standards, legislation, supervisory guidance, research reports)	Secondary summaries of a primary document
Coded sources: relevance	States or finds a condition for an AI initiative to deliver and keep commercial value, or addresses the decision to increase commitment	Sources with no condition that observable evidence can answer
Survey and index sources (not coded)	Admitted whoever produced them (consultancies, vendors, a central bank and a regulator, a university institute), for descriptive statistics only, each with its sample and definition stated	Figures taken from a paraphrase and not read in the source; readings a source does not support, such as the "95 per cent of pilots fail" reading of the MIT NANDA report, which measures something else [15]
Company documents (not coded)	Filings, shareholder letters, press releases, earnings-call transcripts and company-authored research for the two firms in the retrospective case, used for dated events, figures and quotations	Press and analyst commentary on the two firms; figures that a company's own documents give inconsistently
Context sources (not coded)	Theory, history, definitions and method, including practitioner and trade publications	Any use as evidence for a condition or a statistic
Verification	Checked against the publisher's or standard-setter's copy. Where only an abstract, an opening chapter, a bibliographic record or another copy was read, the Declarations say so and the text claims no more than that reading supports	Candidates that could not be confirmed

The review is structured and purposive, not systematic: a systematic review would require a registered protocol, duplicate screening and reproducible coding across databases, which a single-author conceptual paper cannot offer. The six dimensions are the author's synthesis, with each attribution open to check through the Sources column of Table 3.

Coding and analytical procedure

I carried out the analysis in four steps. First, I read each coded source for the conditions it states or finds necessary for an AI initiative to deliver and keep commercial value; a condition was defined as a question that observable

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

evidence can answer, and the unit of coding was the individual finding or claim. Two similar items were counted as distinct when different observations would answer different questions: latency and uptime are measured by different instruments, and controls (the operating controls in place) differ from model governance (the governance of the model lifecycle). Where a vendor and an adopter answer the same question with different observations (price realisation for a vendor, realised savings for an adopter), the two forms are one condition; four such pairs sit in the customer-value and economic rows of Table 3, and counting them separately would give 35 conditions rather than 31. Conditions found in more than one source were counted once and attributed to each; the 31 conditions are the deduplicated set, coded by a single author without a second coder. Second, I grouped the conditions by the question each answers, which produced six adequacy dimensions; where a condition fitted more than one heading I assigned it to the dimension where its evidence is generated and noted the alternative in Table 3. Third, I connected the dimensions to five stages, an exposure rating and four decision outcomes by asking, at each gate, which dimensions must be evidenced and to what level before the next increase in exposure is defensible. Fourth, I read the framework back against the six streams to derive propositions and to separate what it adds from what it restates.

The unit of analysis is the AI initiative as defined in the Introduction. The application in the Findings rests on different material, which that subsection describes.

Findings: The Ai Commercialisation Prudential Framework

The synthesis produced six recurring adequacy dimensions and one cross-cutting decision construct. The dimensions describe what must be evidenced for an initiative to hold its current exposure; evidence adequacy determines whether the organisation has learned enough to justify increasing it. Table 3 lists the dimensions with their conditions and sources; Figure 6 shows how they apply across stages.

Table 3: AICAF adequacy dimensions, core questions, conditions and sources.

Dimension	Core question	Conditions (typical evidence)	Sources
Customer-value adequacy	Does the AI solve a sufficiently important problem for a defined customer or user?	Problem severity; measured outcome improvement against a baseline; repeated use; demand signal (willingness to pay for a vendor, adoption intent and realised benefit for an adopter)	[3,13,19,80,81,82,86]
Technical adequacy	Does the system perform reliably for its intended context?	Task performance; error rates; stress and edge-case tests; drift monitoring; failure analysis	[29,39,53,59,72,75,80]
Operational adequacy	Can the AI operate inside real workflows and supporting systems?	Integration tests; latency; uptime; human-review load (alternative: economic); process fit; incident procedures	[39,44,47,53,69,72,75,78,82,91]
Economic adequacy	Can the initiative create and capture value at sustainable unit economics?	Price realisation for a vendor or realised savings against a baseline for an adopter; cost-to-serve for a vendor or unit cost and cycle time for an adopter; gross margin for a vendor or net benefit after total cost for an adopter; payback; sensitivity analysis	[3,4,9,32,50,60,86,87,102]
Risk and governance adequacy	Are material AI, data, regulatory, security and dependency risks understood and controlled?	Risk assessment including regulatory classification; privacy review; model governance; controls; supplier dependency analysis (alternative: economic)	[4,22,27,28,29,32,39,44,53,69,72,77,87]
Organisational adequacy	Can the organisation own, support and change the initiative at scale?	Accountability; skills; operating model; support capacity; change adoption; executive ownership	[19,47,77,78,80,82,90,91]

Context sources (theory, history, definitions or method; not coded):

[1,2,6,7,8,10,11,12,14,17,18,20,21,23,25,26,30,31,38,40,41,42,46,48,51,52,54,56,57,58,61,68,70,71,73,74,79, 83,88,89]. Survey, index and company-document sources (descriptive figures, dates and quotations only; not coded): [5,15,16,24,33,34,35,36,37,43,45,49,55,62,63,64,65,66,67,76,84,85,92,93,94,95,96,97,98,99,100,101]. Three attributions are explained by the condition coded: [9] and [60] supplied the sensitivity-analysis condition (forward-looking stress testing; option value under uncertainty), and [13] the hypothesis-testing basis of the demand-signal condition. Each condition is rated once, under the dimension that lists it.

Evidence levels and the aggregation rule

Each dimension is assessed at one of four levels defined by what has been observed in the context the gate refers to, its reference context: at G1 a representative workload in a test environment, at G2 the pilot population and workflow, at G3 the live workflow at target volume during the bounded production run described under "Evidence gates", at G4 standing operation at the volume approved at G3, and at G5 the current context for a periodic review or the new context for an expansion. A level grades how well a result has been observed. How good the result must be is the acceptance threshold, a separate matter recorded before the evidence is first presented (see "Calibrating the levels"). Not started means no evidence has been gathered for the dimension. Emerging means evidence exists but from a non-representative setting (volunteers, synthetic data, a single category or task, a demonstration environment), from a single observation, or without a stated baseline. Demonstrated means the result has been observed at least once in the reference context, on a representative population or workload, against a stated baseline, with the method recorded; whether the result is good enough is the separate threshold test. Repeated means the same result has been observed in at least two independent settings or periods under the target operating conditions, with no material deterioration; two periods under one operating or market regime are independent only where the result does not depend on that regime. Controls and procedures use the same words, whichever dimension lists them: emerging when specified but not yet exercised, demonstrated when exercised at least once with the outcome recorded (a change-control procedure applied to a release, an incident procedure run in a test), and repeated when exercised in at least two periods or independent tests, an injected test included. Assessment-type evidence (a risk assessment including the regulatory classification, a privacy review, a supplier dependency analysis) is emerging when drafted, demonstrated when completed and signed off by the challenger, and repeated when refreshed in a second period with no material change in conclusion. A dimension's level is the lowest level among the conditions listed for it in Table 3 that apply to the initiative, each rated once. A condition that does not apply is recorded as such with the reason, and the challenger decides applicability, because declaring a weak condition inapplicable is the cheapest way to raise a level.

The aggregation rule is non-compensatory. Each gate names the dimensions it stresses and the minimum level for each (Table 6); evidence adequacy for the gate is the lowest level among the stressed dimensions, and strength in one dimension cannot offset a floor breach in another. A gate is passed only when two tests are both met on every stressed dimension: the evidence is at the required level, and the result meets the acceptance threshold set in advance. A result observed at the required level that fails its threshold does not pass. Where the result is stable, more evidence of the same result will not help, so the outcome is Remediate where redesign could change the result and Stop where it could not. The minimum is also the natural aggregate for ordinal levels, since an average would assume distances between levels that the scale does not define. Evidence is judged against the specific claim it supports, since customer feedback can mislead on novel value propositions [81]. The binding dimension at a gate is the stressed dimension that stands lowest against its floor: under a Hold or Remediate, the dimension below the required level or failing its threshold; under an Advance, where nothing is below the floor, the stressed dimension that reached its floor last. Under a Stop it is the dimension, stressed or not, on which the stop condition was recorded. Two assessors record a level for each dimension and a tier on each exposure component: the initiative owner, whose record corresponds to the bank's own assessment under SRP20 principle 1, and the independent challenger described below, who corresponds to the supervisory review under principle 2 [9]. Where they differ, the lower level and the higher tier stand until the evidence that would settle the difference has been produced, and the difference is recorded. In effect the challenger's rating stands on the record and the owner's is a disclosure, since the owner's changes the result only when it is the lower one; the forum may still Advance against the record with a stated reason, as "Who challenges" describes.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Calibrating the levels: rubric, raters and the evidence register

Table 17 in Appendix C turns the definitions into a rubric, with illustrative indicators at each level for the six dimensions and one threshold example per dimension, labelled with its level. Whether tying each level to something observable reduces rater inconsistency is untested, and the reliability study in Table 13 is the place to find out.

Levels and thresholds are separate. A poor correction rate observed at target volume against a baseline is demonstrated evidence of a poor result: it reaches the level and fails the threshold. The acceptance thresholds for a gate are recorded at the previous gate, as part of the next evidence set that an Advance defines (Table 5), and confirmed by the challenger before any evidence is presented. G1 has no previous gate, and a return to a gate after a material change has none either; in both cases the thresholds are recorded when the initiative or the change is entered in the inventory, before the evidence is gathered. Thresholds vary by sector and floors by tier, and the two are kept apart so that exposure is not counted twice. A firm's operating thresholds belong in its own policy, approved by whoever holds the mandate for the gate, as Table 4 already asks of the money thresholds. Sector benchmarks are a research task: a structured expert panel using the Delphi method, in which experts answer successive questionnaires with controlled feedback of opinion between rounds [23], could propose them, and Okoli and Pawlowski give guidelines for selecting the experts and designing such a study [73]. The paper supplies no sector benchmarks.

Owner and challenger rate each dimension independently and record their levels before the forum meets, so that neither rating is anchored on the other. A difference of two levels or more sends the evidence item back to its owner for clarification before the gate is heard. Once a year, in each portfolio, the people who serve as owners and challengers rate the same evidence packs without conferring, on levels and on tiers. With several raters and an ordinal scale, the agreement statistic has to be a chance-corrected coefficient that accommodates more than two raters and ordered categories. Krippendorff's alpha is one: Hayes and Krippendorff propose it as a standard reliability measure usable regardless of the number of observers, the level of measurement and the sample size [41]. Cohen's kappa, a coefficient of agreement for nominal scales [17], is the precedent for two raters and unordered categories and does not fit this task. That a coefficient can be computed on a small sample does not make it informative there: three packs give about 18 level ratings per rater, clustered within packs. Such a session reports raw agreement, the disagreements by size (one level, or two or more) and agreement on the binary result the ratings feed, whether the gate's floor is met; the coefficient is kept for the reliability study in Table 13. The benchmark bands usually attributed to Landis and Koch [52] concern kappa-type statistics and do not carry over to alpha; any such band is a convention, and the portfolio sets its own standard in policy before the session. Where raters disagree about whether a floor is met, the pack is discussed until the cause is found, the rubric or the briefing is revised if the cause lies there, and the raters concerned do not sit at a Tier 3 gate until a further session shows agreement.

Each rating rests on an evidence register with one line per item: what the item is, the dimension and condition it supports, the population and period it covers, its date, its source and where it can be retrieved, and who verified it, stating whether that person is the evidence owner. A level with no register entry behind it is recorded as not started. The register makes a rating auditable and is the record from which P1 and P5 would be coded.

A floor is also a target. Bevan and Hood, examining the target regime of the English health service, describe three kinds of gaming: ratchet effects, threshold effects, in which performance crowds towards the target, and distortion of reported output, with evidence that some ambulance trusts had adjusted reported response times to fall inside the target [11]. The threshold effect is the specific risk a floor creates. Three provisions respond. The evidence set and its measures are fixed when a Hold or Remediate is issued. The challenger may draw the sample from the source system, so that the sponsor does not choose the cases shown. Any change to a baseline, a population or a metric definition between reviews is logged and the affected dimension is rated again. Under-rating the tier is covered by the rule that the higher tier stands. Three further routes are for the firm's policy to close, and I suggest defaults: exposure increases are cumulated since the last gate when the tier is re-rated, a Hold renewed twice goes to the mandate holder for a decision between Remediate and Stop, and challengers rotate across initiatives. A fourth matter for policy is what follows when a decision that a challenger supported

is later reversed; the default I suggest is a review of that challenger's ratings at the next calibration session. These provisions reduce the incentive without removing it.

The exposure rating

Exposure is a property of the initiative, not of the stage. It is rated on four components, each defined so that it does not embed another: the people or customers affected by the system's outputs, and how severely; the money committed or at risk; the decision autonomy of the system, from drafting for human review to acting without prior review; and the regulatory class of the use. Table 4 gives a three-tier rating. The initiative's rating is the highest tier reached on any component: a small budget does not make a system that acts without review on customers' money a low-exposure initiative. The organisation sets its own thresholds for "many", "at scale" and the money component and records them in its policy; the money figures in Table 4 are illustrative. In a conduct-regulated firm most customer-facing uses reach Tier 2 on regulatory class, so the rating discriminates there mainly between Tier 2 and Tier 3, the conservative result the maximum rule intends. A use that falls under a prohibited practice, such as Article 5 of the EU AI Act, is outside the rating: it is a Stop, not a tier.

Table 4: The AICAF exposure rating.

Component	Tier 1 (low)	Tier 2 (medium)	Tier 3 (high)
People or customers affected, and how severely	Internal users in one team; outputs reach no customer or third party	External customers, or many internal processes, affected; harm from a wrong output is recoverable (a corrected letter, a refund)	Outputs affect customers' money, health, safety or legal position, and harm is not readily reversible or reaches a population the organisation's policy defines as at scale
Money committed or at risk	Within a team's discretionary budget; withdrawal costs nothing beyond the spend	Above the organisation's threshold for a single budget holder (for illustration, 1 per cent of the unit's annual operating budget), or contracts or integration create a withdrawal cost	Material to the business unit's results (for illustration, a write-off above 5 per cent of the unit's annual operating result, or one that requires disclosure); contract exit or data lock-in would be visible externally
Decision autonomy	Drafts or recommends; a human reviews every output before it takes effect	Acts with sampled or exception-based review	Acts without prior human review, or triggers downstream actions automatically
Regulatory class	No obligation specific to the use; general law still binds, including automated-decision rules such as GDPR Article 22 where they apply	Sectoral conduct rules or transparency obligations apply to the use	High-risk under the EU AI Act, or in scope of a supervisory model-risk regime and classified as high materiality under the firm's own policy

Stage and exposure usually rise together but can diverge: in the constructed insurer case used below, the Stage 2 pilot whose drafts reach customers on regulator-monitored correspondence is Tier 2 on people and regulatory class, while a production summarisation tool used by five internal analysts is Tier 1 on every component. P2 and P6 are tested against the exposure rating, not the stage label.

Decision outcomes

Table 5 sets out the four outcomes. The labels correspond to Cooper's Go, Hold, Recycle and Kill [18]; what is new is the set of rules attached to them. Advance may carry a dated evidence obligation in one situation only: every stressed dimension meets the gate's floor, and a recorded assumption, if wrong, would drop one below it. If the obligation is not delivered, the decision converts to Hold at the review date: exposure stops rising, exposure added since the obligation was set is frozen at the level reached, and the forum decides whether the failed

assumption warrants Remediate or a return to the prior stage, which the decommissioning path tested at G4 makes possible. Stop is final for the initiative as scoped, and a restart enters at G1 with its own record so that sunk cost cannot be carried into the new case. Hold and Remediate must carry a defined evidence set and a review date; without them they become the route by which an initiative avoids Stop, the failure Adner and Levinthal describe when abandonment is not enforced [1]. Each decision records the assumptions on which the evidence depends, in the spirit of the stress testing Pillar 2 requires of banks [9]; the two that matter most for an AI initiative are the cost of inference or human review and a supplier's change of model, API or terms, and further prompts are in Appendix A.

Table 5: Decision outcomes under evidence adequacy.

Decision outcome	Meaning	Expected action
Advance	Every stressed dimension meets the gate's floor, which means that its evidence is at the required level and its result meets the acceptance threshold set in advance, and no stop condition remains, a stop condition being a recorded finding on any dimension, stressed or not, that meets the Stop row below. May carry a dated evidence obligation where a recorded assumption, if wrong, would breach the floor.	Move to the next stage, re-rate the exposure and define the next evidence set with its acceptance thresholds. Deliver any dated obligation or convert to Hold at the review date, with the freeze described in the text.
Hold	Evidence is promising but at least one stressed dimension is below the required level and can be raised by gathering evidence, without redesign.	Stay in the current stage with no rise in exposure beyond the stage's bound; gather the defined evidence set by the review date.
Remediate	A material adequacy gap or control weakness blocks progression and cannot be closed by gathering evidence; it requires redesign of the proposition, the system or the controls. A result observed at the required level that fails its threshold belongs here where redesign could change it.	Return to the relevant stage; redesign, then re-present at the same gate by the review date. A live system may stay in service under change control, with no rise in exposure, while the affected part is reworked.
Stop	The proposition lacks sufficient value, economics or risk viability at the exposure it would need, including a failed threshold that redesign could not change. Final for the initiative as scoped.	Terminate, run the decommissioning path if live, record the learning. Any restart enters as a new initiative at G1.

Evidence gates

Table 6 sets out the five gates, the dimensions each stresses, the minimum level on each, and how the exposure tier moves that floor. The Minimum level column is the Tier 2 floor; where no adjustment is listed for a tier the Tier 2 floor applies. A floor is met only when the evidence is at the level shown and the result meets the acceptance threshold set in advance, so demonstrated evidence of a poor result does not pass a gate. Repeated is the top level, so at G4 the Tier 3 adjustment adds a test of each recorded sensitivity on observed or stressed data, and at G5 Tier 2 and Tier 3 share a floor and differ in review cadence. The tier sets the process as well as the floor (Table 8).

Each gate is the exit gate of the stage with the same number and sits on the arrow to the next stage in Figure 6, and a stage permits only bounded exposure before its exit gate. In Stage 1 nobody outside the team is exposed; G1 precedes any pilot with real users. Stage 2 is a pilot bounded in population, duration and autonomy; G2 precedes entry to the live workflow. Stage 3 is a bounded production run: the system operates in the live workflow at up to target volume for a fixed period, with the prior process kept available as a fallback, no rise in autonomy or tier, and no commitment that would raise the cost of withdrawal. G3 is the gate into production. It precedes full production exposure, meaning standing operation on which the organisation relies and in which the prior process no longer runs alongside the system. The means of returning to that process, or to another fallback, is a different thing: it belongs to the decommissioning path, which is specified at G3 and tested at G4. G3 stresses technical adequacy a second time, because G1's floor of emerging cannot carry a live system. Stage 4 is standing operation at the volume approved at G3 over the two periods G4 needs; commitments that are hard

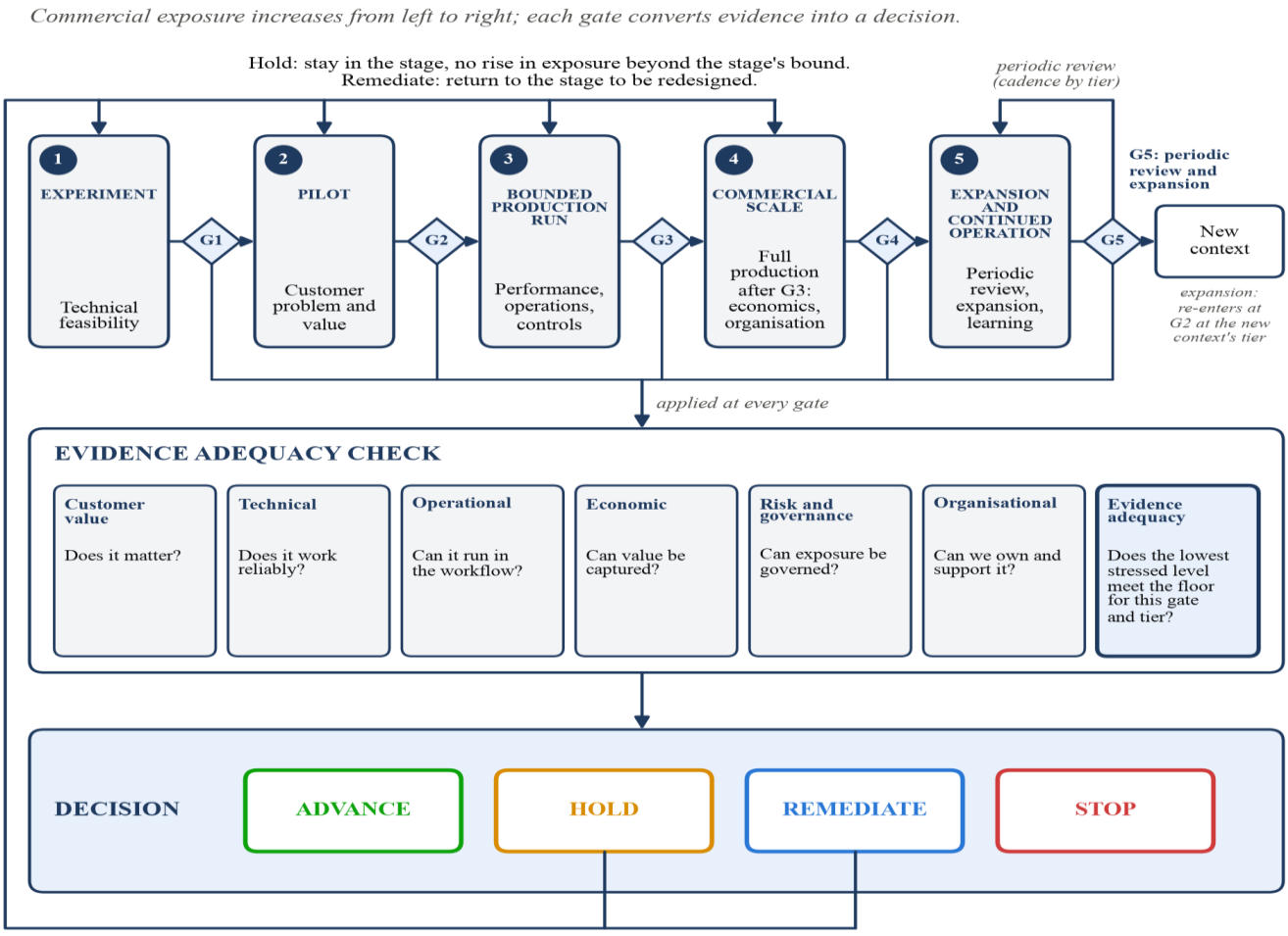
to unwind, such as long-term supplier contracts, volume or price commitments to customers and staffing changes that assume the system, wait for G4. For a system that is high-risk under the EU AI Act a bounded run with real users may itself be putting into service, so the condition attached to Tier 3 at G2 for such systems covers Stage 3's run as well as the pilot.

Table 6: AICAF evidence gates, with the evidence floor by gate and exposure tier. The Minimum level column is the Tier 2 floor. A floor is met only when the level is reached and the acceptance threshold set in advance is met.

Gate	Stressed dimensions	Minimum level	Adjustment by exposure tier	Illustrative gate question
G1: Feasibility	Technical adequacy	Emerging	Tier 3: demonstrated on a representative workload in a test environment, which stands in for the operating context at this gate because no real user may yet be exposed	Is the AI function capable and testable enough to justify a pilot with real users?
G2: Value	Customer-value adequacy; risk and governance adequacy (classification only)	Demonstrated on customer value; emerging on risk and governance, with the regulatory class recorded	Tier 1: emerging on customer value is acceptable. Tier 3: the Tier 2 levels apply. Where the EU AI Act applies and the system is high-risk under it, either the system is conformity-assessed before the pilot or the pilot runs as real-world testing under Article 60 or in a sandbox under Article 57, because a realistic pilot may itself be putting into service	Is there evidence, on a representative group and against a baseline, that the solution improves a meaningful outcome for a defined user or customer?
G3: Operationalisation (the gate into production)	Technical adequacy; operational adequacy; risk and governance adequacy	Demonstrated on all three	Tier 3: repeated on technical adequacy and on risk and governance (results reproduced on a second sample or period; controls exercised in two independent tests)	Does the solution perform and operate reliably within the intended workflow at target volume, with monitoring, change control and an incident procedure exercised, before the organisation relies on it?
G4: Commercial viability	Economic adequacy; organisational adequacy	Repeated on both	Tier 1: demonstrated is acceptable. Tier 3: repeated, with each recorded sensitivity tested on observed or stressed data, which is what Tier 3 adds. Evidence from more than one operating or market regime, where the economics depend on one, is part of repeated at every tier	Do value capture and organisational support hold across two periods of standing operation and under the stated sensitivities, before commitments that are hard to unwind are made?
G5: Expansion and continued operation	All six	Repeated on all six in the current context. Expansion: exposure re-rated for the new context and emerging	Expansion: the new context's tier sets the floor for its own passage through G2 to G4. Periodic review: cadence by	Is the evidence transferable to the new context, is the incremental

		evidence from that context on the dimensions the change affects. Periodic review: the monitoring record checked against the G3 thresholds, and levels re-recorded on any dimension a change has touched	tier (Findings, "After production")	exposure justified, and does the live system still match the evidence on record?

Figure 6: The AICAF commercialisation process. Exposure increases from left to right. Each gate is the exit gate of the stage before it; G3 is the gate into full production, which follows a bounded production run. G5 governs both the periodic review of a live system and expansion into a new context, which re-enters at G2 at the new context's tier. The evidence adequacy check is applied at every gate and produces one of four decisions. On a Hold the initiative stays in its stage with no rise in exposure beyond the stage's bound; on a Remediate it returns to the stage where the redesign is done.



Source: author.

Stage 1: experiment and technical feasibility

At Stage 1 the only question is whether the capability can perform the target task well enough to justify a pilot with real users; benchmark tests, prompt evaluations and controlled experiments supply the evidence, and exposure stays small. Iteration inside Stage 1, and inside a Stage 2 pilot within the bounds G1 approved, needs no gate: a team may change models, prompts, retrieval strategies or interfaces, because no standing operation yet depends on the system. A change that takes a pilot beyond its approved population, duration or autonomy is a change in exposure and goes to a gate.

Stage 2: pilot and problem-value validation

Stage 2 moves the question from "can it work?" to "does it matter?". The initiative must show a meaningful user or customer outcome in a realistic setting, against a baseline, with repeated use and, where relevant, willingness to pay or adoption intent [3,82]. Where humans stay in the loop, the pilot should also test whether the changed workflow can be sustained [78]. G2 also records the regulatory class, for the reason the Discussion gives.

A constructed case shows the gate from the inside; it describes no specific company, and its numbers are invented. A mid-sized insurer pilots a generative AI assistant that drafts responses to customer complaints for 30 handlers on one complaint category. After eight weeks handling time is down by a fifth and handlers rate the drafts highly. Two questions at G2 change the picture: was the outcome measured on volunteers or on a representative group, and what share of drafts were sent without correction? The answers are "volunteers" and "not measured", so customer-value adequacy is emerging against a Tier 2 floor of demonstrated, and the decision is Hold with a representative group, a correction rate and a review date as the evidence set. Eight weeks later a representative group shows the same fall in handling time with the correction rate measured, and the initiative Advances into the bounded production run. The challenger records the regulatory class and its basis at the same gate (Tier 2: conduct rules on complaint handling apply and no Annex III category does).

Stage 3: bounded production run and operational viability

Stage 3 is the bounded production run. The system enters the live process for a fixed period with the prior process retained, and evidence must extend from model performance on the live workload to integration, latency, uptime, monitoring, human-review load and incident response. The deployment, lifecycle and MLOps studies supply the ladder for that evidence [39,47,75]. G3, the gate into production, stresses technical, operational and risk and governance adequacy together. For a provider of a high-risk system under the EU AI Act the obligations must be satisfied before the system is put into service [27], which means before the run; G3 confirms that those artefacts are complete and current.

In the constructed insurer case the assistant is to run for 400 handlers on all categories inside the case-management system; the exposure rating stays Tier 2. At the first G3 review the owner presents integration tests, a four-week monitored run by the 30 pilot handlers on all categories, a drift alert tested once with an injected shift, change control exercised on one prompt update, and an incident procedure run once in a test. The challenger asks at what volume the correction rate was measured. It was 30 handlers, so human-review throughput at 400, and with it the cost per complaint, has not been observed: operational adequacy is emerging, as is technical adequacy, whose error rates were also observed below target volume; the floor is demonstrated, and the decision is Hold. The evidence set is the bounded production run itself, which Stage 3 permits up to target volume: four weeks at 400 handlers with the manual drafting process kept available and every draft still reviewed by a handler. Had the run shown that the hardest category needed a redesigned retrieval step, gathering evidence could not have closed the gap and the decision would have been Remediate. Table 7 shows the worksheet before and after. At the second review all three stressed dimensions are at the required level and meet their recorded thresholds, the initiative Advances, and the assistant becomes the standing process.

Table 7: Worksheet excerpt for the constructed insurer case at G3, before and after the Hold.

Dimension	Level at first G3 review	Level at second G3 review	Key evidence added between reviews
Customer value	Demonstrated	Demonstrated	None required at G3
Technical (stressed)	Emerging	Demonstrated	Error rates by category observed at target volume on the volume run; at the first review they had been observed on 30 handlers only. The drift thresholds and the alert test with an injected shift were already in place at the first review

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Operational (stressed)	Emerging	Demonstrated	Four-week run at 400 handlers: correction rate by category at volume; review queue within service level; incident procedure exercised a second time on a live near-miss
Economic	Emerging	Emerging	Not stressed until G4; cost per complaint now measurable from the volume run
Risk and governance (stressed)	Demonstrated	Demonstrated	None required: change control had been exercised on one prompt update before the first review, and the challenger's sign-off of the risk assessment stands. The incident procedure is rated under operational adequacy and the drift threshold under technical adequacy, each once
Organisational	Emerging	Emerging	Not stressed until G4
Gate result (lowest stressed level, and thresholds)	Emerging: Hold, review in four weeks	Demonstrated, with the recorded thresholds met: Advance	

Stage 4: commercial scale and economic viability

The question at Stage 4 is whether the proposition stays attractive in standing operation. AI economics move with usage because inference, retrieval, data preparation, human review and model-supplier charges scale with demand and are partly set by suppliers [87]. Teece's separation of value creation from value capture is the test [86]: for a vendor, price realisation, cost-to-serve and gross margin across two periods; for an adopter, realised savings, unit cost or cycle time against the pre-deployment baseline, net of the total cost of running and governing the system. Organisational adequacy is stressed at the same gate, because scale tests ownership, skills and support capacity [19,90].

G4 is where the assumption record earns its place. For the constructed insurer two sensitivities decide the economic test: a rise in the supplier's inference price or a change in its terms [32,87], and a correction rate on the hardest category that does not fall with volume, which would make human review scale linearly with demand. In the constructed case the correction rate falls with volume and a named owner and a support rota have operated for two periods, but the margin holds only at a supplier price that can change unilaterally. Every stressed dimension meets the floor and one recorded assumption, if wrong, would breach it, so the decision is Advance with a dated evidence obligation, a renegotiated term or a tested fallback supplier.

Stage 5: expansion, continued operation and portfolio learning

Stage 5 covers expansion beyond the first commercial context and the continued operation of a live system between expansions. A deployment does not carry its evidence to a new segment, geography, workflow or use case, because each change can bring new data conditions, regulations, costs and failure modes; this is MIT CISR's "scaling out" [91]. Expansion re-rates the exposure for the new context and sends the initiative back through G2 to G4 at that tier. Continued operation is governed by the monitoring cadence and change control described below.

Who challenges

Independent challenge has an owner: a party outside the sponsoring unit's reporting line with no delivery target on the initiative, such as a second-line risk or finance function, a product council chaired outside the unit, or, for a small firm, an external reviewer engaged for the gate; where a council is the challenger, the criteria apply to its chair. Two criteria define independence. First, the challenger's reporting line does not run through the

initiative's sponsor, and the challenger's objectives carry no target tied to the initiative's delivery, revenue or savings. Second, the challenger has the competence to read the evidence and the authority to record a level that differs from the owner's. The first criterion scales with the tier: at Tier 1 a peer outside the team with no delivery target on the initiative satisfies it even where both report to the same unit head; at Tier 2 the challenger sits outside the sponsoring unit's reporting line; at Tier 3 the challenger is a second-line function or an external reviewer. The lower of the two levels is the level on record. The decision forum may Advance against it only with a recorded reason and a dated review; that override record is what P1 and P5 read, and it is reported to whoever holds the mandate for the gate. The principle comes from model-risk supervision: SR 11-7 defined effective challenge by incentives, competence and influence [29], its 2026 successor keeps the term [30], and SS1/23 makes independent validation one of its five principles [77]. Sleesman and colleagues' meta-analysis finds personal responsibility for the prior commitment among the significant determinants of escalation [83], which is why the role cannot sit with the sponsor. The worksheet in Appendix A (Table 14) carries columns for both assessors.

Proportional roll-out by tier and organisation size

Size changes who is available to challenge and how many meetings a firm can afford; the tier sets the floor. Table 8 gives the versions. The size bands are the firm's own: what matters is whether a function outside the sponsoring line exists to challenge (medium) and whether there is a second line and a portfolio (large). Where an initiative matches more than one row, the more demanding row governs, so a small firm whose system acts on customers' money without review takes the Tier 3 row whatever its headcount. The lean version, in the manner of Cooper's Stage-Gate XPress and Lite [18], is keyed to tier as well as size. At Tier 1 it hears G1 and G2 as one review at the end of the pilot, which is acceptable because a Tier 1 pilot exposes only internal users in one team. It never combines away the gate into production: G3 is held before full production exposure, and G4 is folded into the first periodic review. The commitments that are hard to unwind, which wait for G4 in the full version, wait for that review in the lean one. At Tier 1 the review is annual and the G4 floor is demonstrated, so one period of operating evidence is enough. For a small organisation at Tier 2 the review is half-yearly and the floor is repeated, so the two periods it needs are quarters or shorter. Every version keeps an inventory of AI initiatives with their tiers, because proportionality presupposes that each initiative has been identified and rated, including those that have not yet come to a gate. Sector enters only through the regulatory class and the thresholds; sector-specific versions are not offered.

Table 8: Proportional versions of AICAF by exposure tier and organisation size. Size bands are the firm's own; floors are those of Table 6 for the tier in every row; where rows overlap the more demanding row governs. In every version a forum override of the challenger's level is reported to the mandate holder.

Version and where it applies	Gates held	Dimensions assessed in depth	Who challenges	Documentation	Review cadence after production
Lean: any organisation at Tier 1; a small organisation at Tier 2	Tier 1: G1 and G2 as one review at the end of the pilot. Tier 2: G1 and G2 separately. G3 always held before full production; G4 folded into the first periodic review, and commitments that are hard to unwind wait for it; G5 as the periodic review	The stressed dimensions; one line each on the others	A peer outside the team at Tier 1; an external reviewer engaged for the gate at Tier 2	Inventory entry, one-page worksheet and evidence register	Annual at Tier 1; half-yearly at Tier 2

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Version and where it applies	Gates held	Dimensions assessed in depth	Who challenges	Documentation	Review cadence after production
Standard: a medium organisation at Tier 2	All five gates, one meeting each	The stressed dimensions, with a level recorded on all six	A finance, risk or product function outside the sponsoring unit's reporting line	Inventory, full worksheet, evidence register and assumption record	Half-yearly
Portfolio: a large organisation at Tier 2 and Tier 3	All five gates; a forum outside the sponsoring unit at Tier 2, a committee at Tier 3	The stressed dimensions, with a level recorded on all six; annual calibration session across the portfolio	A second-line function	Inventory, full record, override log reported to the mandate holder, and portfolio report of rater agreement	Half-yearly at Tier 2; quarterly at Tier 3
Regulated domain, any size (at least Tier 2 on regulatory class)	All five gates, one meeting each	As the standard version, with the regulatory class and its basis recorded at G2	Outside the sponsoring unit's reporting line; an external reviewer where the firm has no second line	Inventory, full worksheet, evidence register, assumption record and classification record	Half-yearly
Tier 3 on any component, any size	All five gates, none combined	All six at every gate from G2, at the Tier 3 floors on the stressed dimensions	A second-line function or an external reviewer, never a peer	Inventory and full record; where the EU AI Act applies, the route recorded at G2: conformity assessment, real-world testing under Article 60 or a sandbox under Article 57	Quarterly

After production: monitoring, change control, retraining and decommissioning

The evidence recorded at G3 and G4 describes the system as it was then, so four rules govern Stages 3 to 5. First, monitoring runs continuously against thresholds set at G3 for the technical and operational conditions in Table 3, and a periodic review at G5 examines the record at a fixed interval: quarterly for Tier 3, half-yearly for Tier 2 and at least annually for Tier 1. The periodic review checks the monitoring record against the G3 thresholds, confirms that no material change has gone unrecorded, and re-records the level on any dimension the record or a change has touched; a full re-presentation on all six dimensions is required only at expansion or after a material change. Second, a material change triggers a return to the relevant gate before the changed system is used. A change is material when the evidence on record no longer describes the system as it will run. The owner describes the proposed change and the challenger records the materiality determination; where they disagree, the change is treated as material, and materiality thresholds scale with the tier: for illustration, any change of model version is material at Tier 3, a version change that fails the regression suite fixed at G3 is material at Tier 2, and at Tier 1 only a change of supplier or purpose is. Material change arises from five sources: the model (a new supplier, version or architecture, or retraining on new data), the data (new sources, populations or feature definitions), the supplier (terms and price, or interface and hosting), the intended purpose (a new user group, a new decision the output feeds, or higher autonomy) and the regulatory environment (a new obligation, a reclassification or an application date). A model, data, interface or hosting change returns to G3; a change of supplier terms or price returns to G4; a change of intended purpose returns to G2; a regulatory change returns to G2 where it changes the class and to G3 where it changes the controls required. A breach of a monitoring threshold also returns to G3 without waiting for the periodic review. A return to G3 from standing operation need not take the service down: where the unaffected part of the system or the fallback can carry the work, the service continues under change control while the affected part is reworked and re-presented. A rise in volume, people or money in the same context beyond what the last gate approved is a limit increase and has its own route. The exposure is re-rated first, and a higher tier brings that tier's floors. Operational adequacy at the new volume is then re-presented at

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

G3 as a bounded run at that volume, and the economic evidence at the new volume returns to G4 before further commitments that are hard to unwind. Third, retraining under a procedure validated at G3, with stated bounds on what may change, is logged and covered by monitoring; retraining outside those bounds is a material model change. Fourth, decommissioning is a planned path whose evidence (a fallback process, user or customer notice, data retention and deletion, contract exit) is specified at G3 and tested at G4, so that Stop remains available for a system that 400 people depend on.

Why not a single score

AICAF avoids a universal numerical score, where Cooper's SPARK model proposes scoring to counteract intuition bias [20], a concern the escalation literature supports [83]. The objection to a single score is that it is compensatory: an initiative with excellent technical results and no incident procedure can outscore one with modest results and full operational readiness, and in the constructed insurer case a score would have carried the initiative through G3 before human-review throughput had been observed. Stage-Gate's must-meet criteria answer the same objection for development decisions [18]; whether SPARK's scored factors sit behind knock-out criteria is not stated in [20]. AICAF applies the must-meet logic to every stressed dimension, with the level scale giving "must meet" observable content.

Retrospective application of the dimensions and outcomes to eight initiatives, with two excluded

I have read eight AI initiatives that I led or advised over two years through the framework's six dimensions and four outcomes. The framework in its present form dates from 2026, and I applied its terms to these decisions afterwards, with the outcomes known. On four of the eight I was Chief Product Officer and the accountable decision-maker at the time. On the other four I advised as a product and commercialisation consultant, and the reading is reconstructed from that position. Two further initiatives are excluded: both were paused before production, and their later outcome could not be established. Excluding them tilts the eight towards initiatives that advanced. A pause whose outcome is never established is also the situation for which the review date on every Hold and Remediate is written.

Table 9 lists the eight with the organisations unnamed. Furthest stage is the last AICAF stage the initiative reached on the stage definitions of this paper, in which Stage 3 is the bounded production run. In production means standing operation after the gate into production, which is Stage 4 or later. The decision column gives the AICAF outcome to which I read the most recent gate decision as corresponding; for initiative 2, which has since reached Stage 5, it gives the Advance at G3 that I describe below, with the earlier Remediate at G3. The binding dimension is used here in a plainer sense than the definition under "Evidence levels and the aggregation rule", which needs a gate, a floor and recorded levels that I do not have. In Table 9 it is the dimension on which the decision turned, as I read it. For an Advance that is the dimension whose evidence settled the decision to proceed; for a Remediate, the dimension whose shortfall sent the work back; for a Stop, the dimension on which the case failed.

Table 9: The author's retrospective application of the AICAF dimensions and outcomes to eight AI initiatives. Organisations are unnamed; every reading is the author's own and was made after the decisions, with the outcomes known. The binding dimension is the dimension on which the decision turned, as the author reads it, assigned without recorded gates, floors or levels.

#	Sector	Author's role	Furthest stage	Decision	Binding dimension	In production
1	Banking	Chief Product Officer	5	Advance	Risk and governance	Yes
2	Communications	Chief Product Officer	5	Advance, after one Remediate	Risk and governance	Yes
3	Banking	Chief Product Officer	5	Advance	Technical	Yes

4	AI	Product and commercialisation consultant	4	Advance	Customer value	Yes
5	Learning management systems	Product and commercialisation consultant	5	Advance	Customer value	Yes
6	Communications	Chief Product Officer	4	Remediate	Operational	Yes
7	Learning management systems	Product and commercialisation consultant	4	Advance	Operational	Yes
8	AI	Product and commercialisation consultant	2 (pilot)	Stop, before production	Economic	No

Two of the eight initiatives were in each of banking, communications, AI and learning management systems. Six decisions read as Advance, one of them after a Remediate (initiative 2), one as Remediate and one as Stop; none reads as a Hold. Four initiatives reached Stage 5 and three reached Stage 4. The eighth was stopped at the pilot, Stage 2, and never ran in production; the other seven are in production. On my reading, risk and governance, customer value and operational adequacy were each binding on two initiatives, technical and economic adequacy on one each, and organisational adequacy on none; that last count may reflect my seat as product lead or adviser as much as the initiatives themselves. Of the three initiatives at Stage 4, two turned on operational adequacy and one on customer value. I read the other three Stage 5 rows (initiatives 1, 3 and 5) at G5, which stresses all six dimensions. On that reading the dimension named is not one that the corresponding gate stresses in two rows: customer value for initiative 4 at an advance into Stage 4, where G3 stresses technical, operational and risk and governance adequacy, and economic adequacy for initiative 8 at a stop in the pilot, where G2 stresses customer value and risk and governance. In those rows the reading shows the decision turning on evidence that the gate would not have required at that point; for the Stop, Table 5 allows a stop condition on any dimension, stressed or not. These are counts from eight cases that I read myself, and I draw no causal conclusion from them.

Initiative 6 was already running in production when an operational shortfall sent part of it back for rework while the service stayed up. I read that as a Remediate under the change-control rule: a return to G3, which stresses operational adequacy, with the service continuing. That is why its row shows Stage 4, Remediate and in production together. I do not report whether the rework has since been re-presented.

Initiative 2 is the one case I describe at length. An AI compliance-validation capability, using retrieval and generation on a communications platform in a regulated environment, was kept out of production until four constraints were in place: retrieval restricted to approved organisational documents, a confidence rating on every answer, source lineage on every answer, and a human accountable for every final approval. During the pause the four constraints had to be built, which was design work. The team then validated the whole workflow end to end against the organisation's approved documentation, which took three weeks. A gap that needs redesign of the controls is a Remediate under Table 5, so I read the pause as a Remediate at G3, the gate into production, with risk and governance binding, followed by an Advance. That reading came later. The capability went into production and, on my own statement, which readers cannot verify, was later adopted by a large private-sector bank; I offer no G5 reading of that adoption. Stakeholders responsible for the regulated client environment challenged the production decision, and I cannot document the specific role of that challenger. I also lack the month of the pause, dated document references and a baseline for the three weeks, which therefore carry no claim of a reduction. The controls evidenced were design-level. I report no error rate, no calibration of the confidence rating and no post-launch monitoring record, because none is available to me.

The following were not recorded for any of the eight, and I have not reconstructed them: the exposure tier and the component that set it, the year and quarter of each decision, the evidence level on each dimension, the acceptance thresholds, and whether an outside challenger took part, initiative 2 apart. That is why Table 9 has no gate, tier, level or challenger columns. The application shows that past decisions can be read through the six

dimensions and four outcomes. It does not exercise the exposure rating, the level floor, the threshold test or independent challenge.

The application has further limits. I designed the framework and made every reading, the binding dimension included, knowing the framework's thesis. All eight readings are retrospective and were made with the outcomes known. The eight are my own portfolio and were not sampled, and the exclusion of two paused initiatives tilts them towards initiatives that advanced. I describe the eight with the permission of the organisations concerned; no organisation is named and no confidential detail is disclosed. The underlying records are confidential and are not available to readers.

A retrospective reading of two public cases: Zillow Offers and GitHub Copilot

I also read two public records against the gates: Zillow Offers, which Zillow Group decided to wind down in November 2021 [99], and GitHub Copilot, which Microsoft described in July 2024 as a larger business than the whole of GitHub had been when it was acquired [67]. "Public record" here means company and regulatory documents only: filings, releases, research posts and earnings transcripts. Press and analyst commentary, independent research and litigation are excluded by Table 2, which also screens out any outside account of management decisions. One person coded the reading, with the outcomes known and no access to either firm's internal decisions or evidence, so every level in Table 10 describes those documents and may understate what the firms held. The stage mapping is the coder's; neither firm's sequence followed the framework's order.

For Zillow the reading starts with the 12 April 2018 announcement that the firm would itself buy and sell homes in two markets [93] and resumes in February 2021. The years 2018 to 2020 are not read, because purchase volumes before the fourth quarter of 2020 were not extracted from the filings; the G3 and G4 decisions on the framework's stage definitions probably fell in that period, so the 2021 rows are read as returns to gates after a material change. A control for halting purchases existed and was used in 2020 [98]. When the Zestimate became an initial cash offer on 25 February 2021 [95], autonomy rose, which returns the initiative to G2 and on through G3. The one accuracy figure then public, a median error of 1.9 per cent [95], measured the valuation of homes already listed, whereas the company later described its need as forecasting prices three to six months ahead [100]. For the forecasting task that evidence comes from outside the reference context, which is the definition of emerging and is below the G3 floor on technical adequacy at any tier. I rate money at Tier 3 on figures public before the outcome: the second-quarter letter reported 3,805 homes bought in the quarter with the firm's own capital, 3,142 held in inventory at its end and guidance for a further step up [97]. The letter appeared on 5 August, after the purchases it reports, so the rating applies to the step from that guidance onward. The Homes segment's results consisted of buying and reselling that inventory, so the commitment was material to the unit under Table 4 whatever happened later; rating the tier from the write-down would use the outcome as an input. At G4 the floor is repeated economic evidence. Favourable results had been reported in more than one quarter, but in a rising market, and the fourth-quarter 2020 letter had attributed its result primarily to stronger home price appreciation than expected [94]. Periods under one market regime are not independent where the result depends on the regime, so I read the level as demonstrated and the floor as indicating Hold at the second-quarter volume, with forecast error over the holding period and renovation throughput as the evidence set.

The company's own account of the loss is forecasting unpredictability and capacity [99,100,101]; its documents do not show whether the market, the models or operating decisions caused it, or whether an internal limit was breached. Zillow's exposure was inventory bought with its own capital, for which the prudential instruments are position limits and loss-absorbing capital. AICAF borrows neither, so the case illustrates the limit stated in "Prudential logic" as much as it illustrates the gates.

Copilot suggests code and the developer accepts or rejects each suggestion, which is Tier 1 on autonomy. General availability on 21 June 2022 at USD 10 a month [35] extended the people affected to any paying developer, Tier 2 on that component. The preview had already grown from limited places to more than 1.2 million developers [33,35], a rise in people affected that the framework would have gated. At general availability the public evidence on customer value was twelve months of preview use and one study by GitHub authors relating accepted suggestions to perceived productivity [92], with no outcome against a baseline, so the level was emerging against a floor of demonstrated. A controlled experiment had ended the day before and was published on 7 September

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

2022: developers given Copilot completed a set task 55 per cent faster than a control group [49,76]. I apply the level definitions to it as I do to Zillow's accuracy figure. GitHub and Microsoft staff ran it on one task with 95 programmers recruited through a freelance platform [76]; a single task and a recruited sample make it emerging evidence under the definitions, stronger than before and still short of demonstrated. The demand signal is stronger, with over 1 million paid users by October 2023 [66], but a dimension takes its lowest condition. Risk and governance and economic adequacy cannot be read from the company's documents, which disclose no cost to serve or margin, and they do not let me place G3 or G4; "successful" rests on the vendor's adoption and revenue disclosures [66,67].

The framework therefore returns Hold for the successful case at general availability, and on the documents read that Hold would not have cleared in the months that followed. I regard this as a possible false Hold of the kind the Discussion describes. The step added paying users who reviewed every suggestion and paid by the month, the firm may have held evidence it did not publish, and a floor of demonstrated may be too strict for a step so cheap to reverse; the case shows the cost of the rule and cannot show whether the rule is wrong. The contrast between the cases concerns the size and reversibility of each exposure step and makes no claim that one firm used gates and the other did not: on the documents read, both raised exposure ahead of the floor. Zillow's step added homes to be bought, renovated and resold at about five times the quarterly volume of two quarters earlier [96,100], with the algorithm opening the offer. Both cases were selected on known outcomes and read with hindsight. They illustrate the framework's vocabulary and test none of its propositions.

Table 10: Exposure steps, public evidence and AICAF reading for Zillow Offers and GitHub Copilot. Levels describe the company documents cited and nothing else; the reading is retrospective and by a single coder.

Date	Exposure step	Component raised	Evidence on the public record	AICAF reading
Zillow Offers				
12 April 2018	Zillow to buy and sell homes in Phoenix and Las Vegas [93]	Money: the firm's own capital	Seller behaviour in the Instant Offers test; nothing published on pricing accuracy [93]	G1, entry to a bounded pilot. Customer value emerging; technical evidence not on the public record
25 February 2021	Zestimate becomes an initial cash offer in more than 20 cities [95]	Autonomy; people affected	Median error of 1.9 per cent for on-market homes; a limited subset of homes [95]	Material change returning to G2 and on through G3. Technical emerging for the forecasting task, below the G3 floor; the limited subset is consistent with holding exposure
Second quarter 2021, reported 5 August	3,805 homes bought, more than double the first quarter; 3,142 in inventory [97]	Money	Gross profit of USD 33,849 per home sold [97]; results in a rising market only, with no flat-market or falling-market test in the documents [94,97]	G4 at Tier 3 on money, rated on the purchases and inventory disclosed at the time. Economic demonstrated against a floor of repeated, the periods sharing one market regime: Hold at that volume
Third quarter 2021, reported 2 November	9,680 homes bought; 9,790 in inventory [100]	Money	Reported after the quarter; renovation and resale capacity under pressure [100]	Volume rose beyond the level evidenced, which reopens G3 on this reading: operational evidence at this volume absent

18 October and 2 November 2021	New contracts suspended; board decides to wind down [98,99]	Exposure frozen, then reduced	Inventory write-down of USD 304.4 million [99]	Stop, with a decommissioning path run
GitHub Copilot				
29 June 2021	Technical preview with limited places [33]	People: a limited number of developers; autonomy at suggest-only	No effectiveness data published [33]	G1, entry to a pilot. Tier 1 on autonomy
21 June 2022	General availability at USD 10 a month [35]	People: all developers. Money: a paid product	Preview use by more than 1.2 million developers [35]; one vendor study of perceived productivity [92]; no public baseline	G2 at Tier 2 on people. Customer value emerging against a floor of demonstrated: Hold, a possible false Hold (see text)
7 September 2022	Controlled experiment published; it had ended the day before general availability [49,76]	None	95 recruited developers; task completed 55 per cent faster with Copilot; vendor-run, one task [49,76]	Customer value still emerging under the level definitions (one task, recruited sample); stronger evidence, not independent
7 December 2022 and 14 February 2023	Copilot for Business at USD 19 per user per month, then open to every organisation [36,37]	People: organisations	The experiment above; policy controls and a privacy commitment stated [36]; nothing in the company's documents on code quality, security or licensing risk	G5 expansion returning to G2. Risk and governance cannot be read from the company's documents
October 2023 and July 2024	Vendor disclosures [66,67]	None	Over 1 million paid users and more than 37,000 organisations [66]; Copilot above 40 per cent of GitHub's revenue growth in the year [67]	Demand signal strong. Economic adequacy cannot be read: no cost to serve or margin disclosed. G3 and G4 cannot be placed

DISCUSSION

Why technical readiness is a poor proxy for commercial readiness

The framework reframes commercialisation as a problem of exposure maturity: a TRL asks whether a technology is mature enough for a technical context, AICAF whether the organisation has enough evidence to carry the consequences of commercial exposure. Two initiatives can sit at the same readiness level and face very different risk: an internal document-summarisation tool and a credit-decision assistant may both be at MLTRL level 8, yet under Table 4 the first is Tier 1 and the second is Tier 3 on people and regulatory class, and Table 6 sets them different floors at G1 to G4. Readiness is one input and never a proxy [59,74].

What AICAF adds to the process frameworks it resembles

Table 11 positions AICAF against the methods it most resembles, limited to what each source states. Agile-Stage-Gate [21] is treated as Stage-Gate with sprints inside stages and shares its row. Table 12 gives my judgement of when each method is the better choice.

Table 11: AICAF compared with related process and risk frameworks.

Framework	Unit of decision	Criteria type	Scoring approach	Evidence burden escalates with exposure	Independent challenge	Post-launch coverage
Stage-Gate with must-meet criteria [18]; Agile-Stage-Gate [21], treated here as Stage-Gate with sprints inside stages	Funding of the next development stage, with projects also prioritised against one another at gates	Must-meet (knock-out) and should-meet criteria	Should-meet criteria scored and weighted; must-meet are yes or no	The process is scaled to project risk and size (Stage-Gate XPress and Lite); gate criteria are not set per dimension by an exposure rating	Gatekeepers are senior resource owners; independence from the sponsoring line not required	A post-launch review follows the final stage
SPARK [20]	Go or kill on an AI project in new product development	Scored factors; whether knock-out criteria sit behind them is not stated in [20]	Single weighted score inside a gated process	Not by exposure; otherwise not stated in [20]	Not stated in [20]	Not stated in [20]
Lean Startup [12,79]	Whether to persevere with a business hypothesis or pivot [81]	Validated learning from hypothesis tests, with an emphasis on quickly observable customer feedback [31,81]	None; innovation accounting tracks cohort metrics against a baseline, with no aggregate score [79]	No exposure rating. Learning milestones (establish a baseline, tune, then pivot or persevere) are the nearest thing to a staged threshold, and the minimum viable product keeps early exposure small [79]	Not required; the pivot-or-persevere meeting is held by the product and business leadership [79]	Continuous; the Build-Measure-Learn loop carries on after launch [79]
MLTRL [53]	The system's readiness level, 0 to 9	Level-specific review requirements (verification and validation, integration, privacy and security)	Ordinal level	With technical maturity; commercial exposure is not a variable	Reviews at level transitions; independence from the development team not specified	Level 9 deployment includes monitoring within the level
NIST AI RMF [69]	Risk management of an AI system across its lifecycle	Outcome-based subcategories under Govern, Map, Measure and Manage	None; voluntary and non-prescriptive	Risk-based prioritisation; no gate and no economic criterion	Encouraged; no gate authority defined	Yes: monitoring, and mechanisms to supersede or deactivate, within Manage
AICAF	The decision to raise commercial exposure	Non-compensatory evidence floor per stressed dimension	Ordinal level per dimension; minimum rule; no aggregate score	Yes: floor set by gate and by a three-tier exposure rating	Required; two independence criteria scaled by tier, and an owner	Yes: G5 periodic review by tier, change control, retraining rule, decommissioning path

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Table 12: Decision matrix: conditions under which each method is most appropriate. The conditions are the author's judgement from the characterisations in Table 11 and have not been tested comparatively.

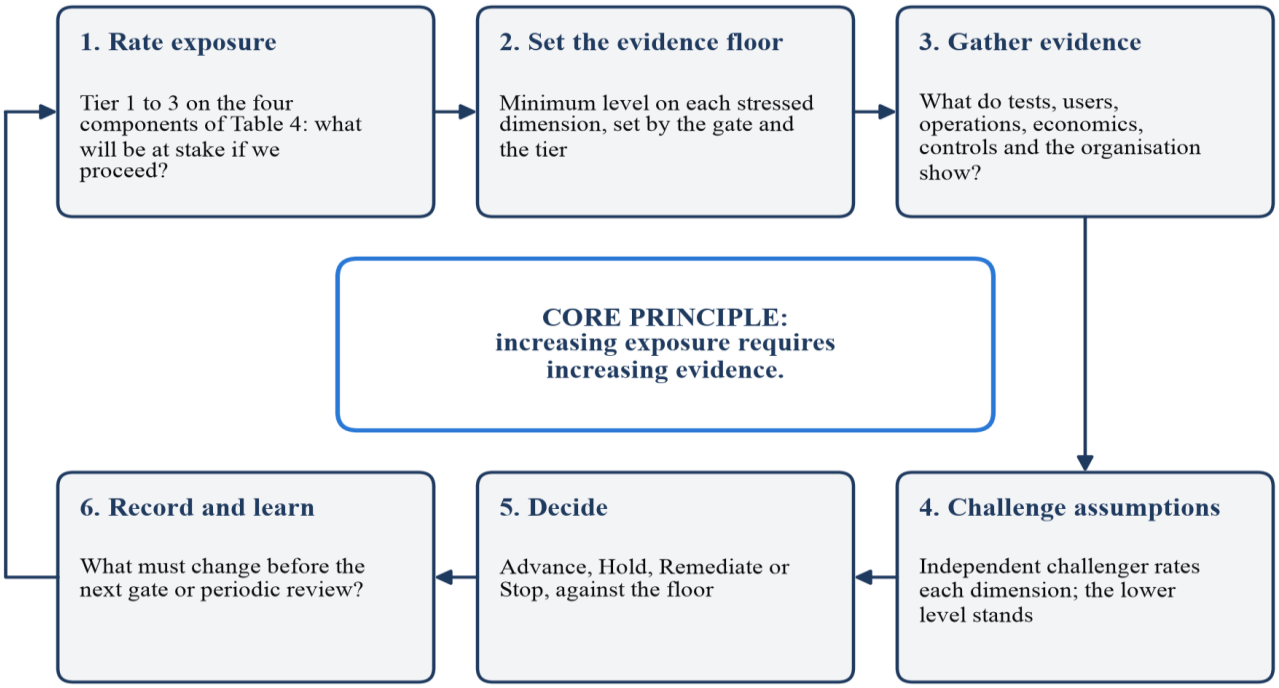
Method	Most appropriate when	Another method is the better choice when
Stage-Gate [18], including Agile-Stage-Gate [21] where the product is built in sprints	A new product has a definable business case, gatekeepers control the resources, and the gates test market and business-case risk as well as development cost	Projects at the same stage differ widely in autonomy, people affected or regulatory class (add the AICAF rating); or the product concept is still unknown (Lean Startup first)
SPARK [20]	Go or kill decisions on AI projects in new product development are being made on intuition and a score is wanted to counter that bias	A weakness on one dimension must not be offset by strength on another, as with operational or risk evidence before production (AICAF or must-meet criteria)
Lean Startup [12,79]	The customer problem and business model are still being searched for, exposure is small and feedback is quick to observe	The evidence that matters is slow or costly to observe, such as operational, economic or risk evidence at scale [31], or other people already depend on the system (AICAF)
MLTRL [53]	An ML system is being matured as engineering and technical failure is the dominant risk	The decision is commercial, since commercial exposure is not a variable in the scale
NIST AI RMF [69]	An organisation is establishing AI risk management across all its systems and no single investment decision is at issue	A decision to proceed needs an economic criterion, which the framework calls for and does not supply
AICAF	An AI initiative is past the idea stage, the question is whether to raise its commercial exposure, a mandate for the gate exists and a challenger sits outside the sponsor's line; the fit is closest at Tier 2 and Tier 3	The tool has a single user or the work is a Tier 1 experiment (plain iteration or Lean Startup); no mandate or independent challenger exists (AICAF is advisory only); probabilities and payoffs can be estimated (a decision tree or value-of-information analysis [42,56]); or every AI project is low-exposure and Stage-Gate with must-meet criteria is already in use

The methods combine. Lean startup experiments can produce the customer-value evidence that a G2 review then weighs, and a firm already running Stage-Gate can add the exposure rating and the challenger criteria to its existing gates without replacing them.

A Stage-Gate gate with must-meet criteria already decides whether to fund the next stage against knock-out conditions, and scalable Stage-Gate already varies the process with project risk and size [18]. An AICAF gate asks four things that such a gate does not: the exposure rating, so that two initiatives at the same stage can face different floors; a non-compensatory adequacy floor per dimension whose level rises with that rating; the assumption record, so that the decision states what would invalidate it; and a challenger meeting two independence criteria, where Stage-Gate requires a gatekeeper with authority over resources. The exposure rating, the assumption record and the challenger come from the prudential sources, as does the principle of a floor set in advance that rises with what is at stake; the floor's non-compensatory property comes from Stage-Gate's must-meet criterion, as Table 1 records. The rising floor also has a decision-theoretic reading. In Wald's sequential test the stakes enter through the error probabilities the analyst states [89], and raising the floor with the tier is the ordinal analogue of tightening those probabilities. A gate is triggered by a change in exposure on

any of the four components in Table 4 or, from Stage 3 onwards, by a material change under the change-control rule. Figure 7 draws the loop.

Figure 7: The commercialisation decision loop: rate the exposure on the four components of Table 4, set the evidence floor for the gate and tier, gather the evidence, record the independent challenger's level, decide among the four outcomes, and record what must change before the next gate or periodic review.



Source: author.

Where governance evidence enters the investment decision

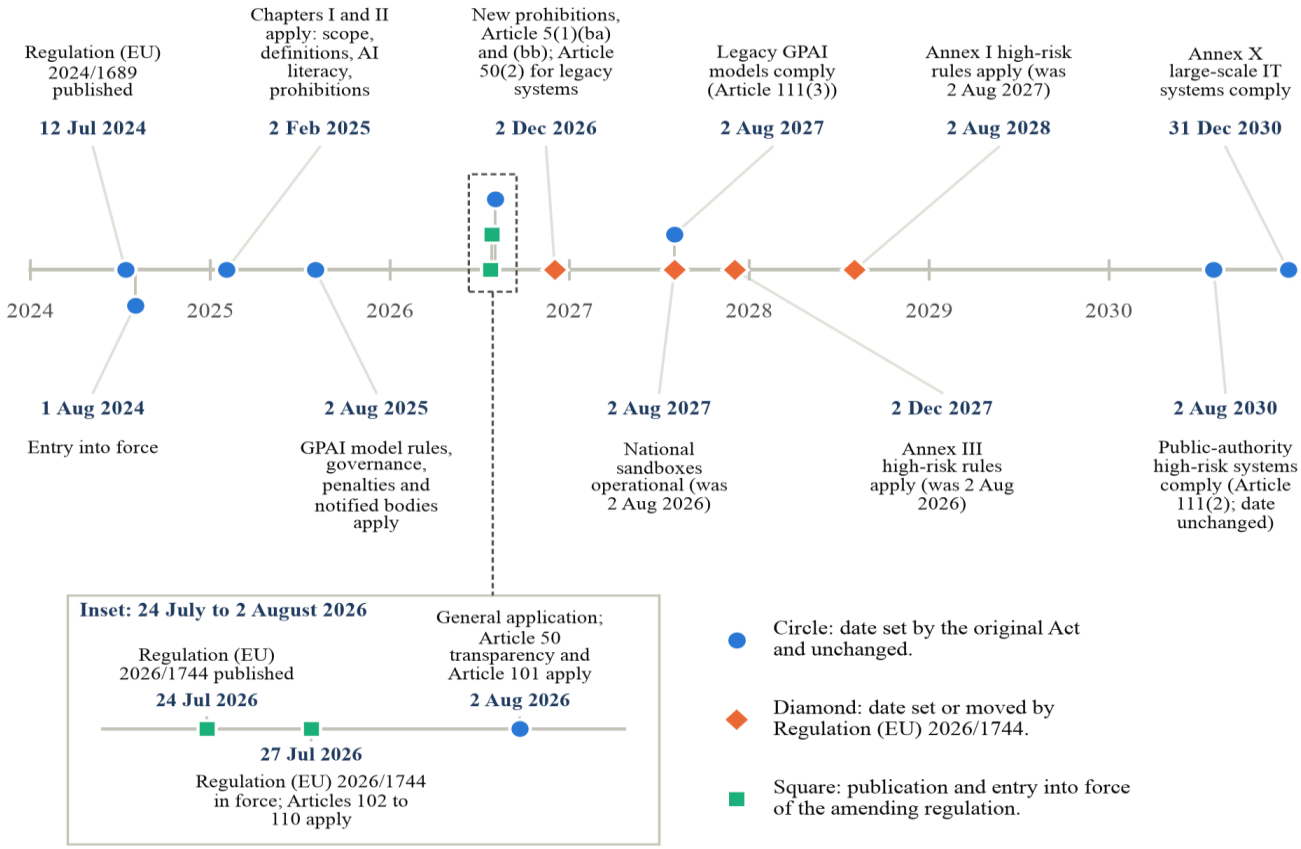
NIST AI RMF, its Generative AI Profile and ISO/IEC 42001 provide the mechanisms for managing AI risk and establishing organisational governance [44,69,72]. AICAF consumes their outputs: a risk assessment, model evaluation, privacy review or incident-management plan becomes evidence for a commercial decision, where before it sat as a compliance artefact apart from product and investment governance.

The EU AI Act is the clearest public example of obligations tiered by consequence, and two of its features matter for the gates. The first is role. A provider develops a system, or has it developed, and places it on the market or puts it into service under its own name (Article 3(3)); a deployer uses a system under its authority (Article 3(4)); and Article 25 turns a deployer into a provider if it puts its name on a high-risk system, substantially modifies it, or modifies the intended purpose of a system such that it becomes high-risk [27]. Provider obligations for high-risk systems (Article 16) attach at placing on the market or putting into service, which Article 3(11) defines as supply for first use, including the provider's own use. A realistic pilot with real users may therefore already be putting into service unless it runs as testing in real-world conditions under Articles 60 and 61 or in a regulatory sandbox under Article 57; Article 2(8) excludes research and testing from the Act's scope but not testing in real-world conditions. That is the basis of the Tier 3 condition at G2 in Table 6. Deployer obligations (Article 26) attach to use from the date the relevant chapter applies. The framework's vendor and adopter labels are not the Act's roles, and an adopter can be a provider: the constructed insurer commissions the assistant and puts it into service under its own name, so it is a provider under Article 3(3), though a tool that drafts complaint responses for handler review is unlikely to fall in an Annex III category, and the Article 16 obligations attach only if it does.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

The second feature is timing. The Act entered into force on 1 August 2024, with general application from 2 August 2026 [27]. Regulation (EU) 2026/1744, which the Commission calls the Digital Omnibus on AI [26], entered into force on 27 July 2026. It moved the high-risk obligations for Annex III use cases from 2 August 2026 to 2 December 2027 and for Annex I product-embedded systems from 2 August 2027 to 2 August 2028, added two prohibitions applying from 2 December 2026, and moved the date by which each Member State must have a national regulatory sandbox operational to 2 August 2027 [28]. Figure 8 draws the timeline, and Table 16 in Appendix B gives the legal basis of each milestone. A firm whose system falls in Annex III has a regulator-set date by which its provider or deployer evidence must be complete. For firms outside the Union's reach the regulatory floor is whatever binds them: SS1/23 for a UK bank within its scope [77], sectoral rules elsewhere, or a voluntary baseline such as the NIST framework [69]; the regulatory-class component of Table 4 takes whichever regime applies.

Figure 8: Application timeline of the EU AI Act, 12 July 2024 to 31 December 2030, as amended by Regulation (EU) 2026/1744. Fourteen markers on thirteen dates are drawn; Table 16 in Appendix B gives the milestone and legal basis of each. Circles mark dates the original Act set and the amendment left unchanged; diamonds mark dates set or moved by Regulation (EU) 2026/1744; squares mark the publication and entry into force of the amending regulation. At 2 August 2027 a diamond and a circle are drawn as separate markers, and the July to August 2026 cluster is shown in an inset.



Source: Regulation (EU) 2024/1689 [27]; Regulation (EU) 2026/1744 [28]; legal text read on EUR-Lex on 27 September 2026.

Boundary conditions and the cost of a false Hold

AICAF adds little in four settings. A single-user tool has no exposure to govern beyond the user's own time. A Tier 1 initiative needs only the lean version in Table 8. Research settings, where no system is deployed, involve no exposure decision. A bought SaaS feature, where the supplier holds the technical evidence, shifts the burden: the adopter relies on what the supplier can be made to provide (evaluation results, model documentation, contractual evaluation rights, change notification) and generates its own evidence on customer value, operational

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

fit, economics and organisational adequacy. In the Act's terms such a buyer is a deployer under Article 26 unless Article 25 applies, and a feature embedded in an incumbent platform, where the adopter cannot choose the model, leaves it with less technical evidence of its own than a bought AI product does. Continuously learning or agentic systems, whose autonomy can change between gates, are covered by the autonomy component of Table 4 and the change-control rule, which treat a rise in autonomy as a material change.

The framework also has a cost. Real-options theory cuts the other way at low exposure: the value of a small staged commitment rises with uncertainty [60], so demanding evidence before a cheap experiment destroys option value. A false Hold delays learning, holds a team in place and can cede a market. Howard gives the test in principle: placing a value on the reduction of uncertainty is the first step in deciding how much to spend on reducing it [42], so a Hold earns its delay only when the evidence awaited bears on a consequence large enough to outweigh the wait. The gate rule does not make that comparison, since the floor rises with the stakes and never falls with the cost of waiting, and the Copilot reading in the Findings is a candidate example. The proportionality rule limits the burden; the review date on every Hold and the escalation of a twice-renewed Hold are the defence against the gate that never decides.

Propositions for empirical testing

The conceptual model generates six propositions, each naming its variables; the records from which they can be coded are gate minutes and worksheets, stage-transition dates, budget by stage, and incident, rollback, supplier-change and decommissioning logs. For portfolios that have not adopted the framework, levels and tiers are coded retrospectively by the researcher from the gate records, which is the main measurement risk for P1, P3, P4 and P6.

P1: Among initiatives that advance past G3, those with any stressed dimension recorded below demonstrated show a higher rate of Stop or rollback within 12 months than those with all stressed dimensions at or above demonstrated (independent variable: recorded levels at G3; dependent variable: Stop or rollback to a prior stage within 12 months). The first group exists only by recorded forum override in a portfolio applying the framework, or by retrospective coding in a portfolio that does not.

P2: Decision forums using an exposure rating and a non-compensatory floor show higher inter-rater agreement on Advance decisions for the same evidence pack than forums that do not, measured in a vignette experiment.

P3: The effect of technical adequacy level at G3 on reaching G4 within 18 months is carried by operational adequacy level: technical level predicts operational level, operational level predicts progression, and little direct effect remains [39,75].

P4: Among initiatives with customer-value adequacy at demonstrated or above at G2, the share still in production 24 months after G4 is lower where economic adequacy at G4 was below repeated than where it was repeated [3,87].

P5: Gate decisions taken with a challenger meeting the two independence criteria are reversed within 12 months less often than decisions taken without one, with a larger difference where recorded sunk cost or time in stage is high [1,13,83]; the share of Hold and Stop outcomes at G2 to G4 serves as a check that the challenger's presence changed decisions, not as an outcome in itself.

P6: Portfolios whose recorded evidence floor rises with the exposure tier show a lower rate of post-G4 reversal (rollback, write-off or Stop within 24 months) among Tier 3 initiatives than portfolios whose floor is uniform across tiers or set by model complexity [29,77].

P1 and P5 should be tested first, because both can be coded from records that firms already keep and P1 is the proposition on which the framework is most easily falsified. P6 follows, because it carries the framework's distinctive claim; P2 needs a vignette experiment, and P3 and P4 need validated measures of the levels over time.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Limitations And Future Research

Six limitations apply. The first is that the only application is my own and is retrospective. I designed the framework and read the eight initiatives in the Findings myself, after the decisions and with the outcomes known, and no exposure tier, evidence level or decision date was recorded for them. Two initiatives that were paused before production, and whose later outcome could not be established, are excluded, which tilts the eight towards initiatives that advanced. No independent rater, sampled portfolio, expert panel or comparative test stands behind AICAF, and the reading of Zillow Offers and GitHub Copilot is illustrative secondary analysis by one coder with the outcomes known. If a gated portfolio performed no better than a conventionally reviewed one, the framework would be a taxonomy. Second, the review is purposive, the coding had no second coder, the decision-under-uncertainty stream was not coded, and some works on decision theory and measurement were read in part or as abstracts only, as the source-verification note states. Third, the prudential analogy has an enforcement boundary: without a mandate and an independent challenger the framework is advisory, and supervisors' own use of model-risk logic for AI [22,32] makes its transfer across domains plausible without proving it, and the 2026 US model-risk guidance places generative and agentic AI models outside its scope [30]. Fourth, ordinal levels without a numerical score leave the framework open to the objection that judgement is less consistent than a scorecard [20]. The rubric, the two-assessor rule and the calibration session are designed to reduce that risk and none has been tested; the thresholds in Appendix C are illustrative, and sector benchmarks and sector-specific roll-out are deferred. Fifth, a floor gives a sponsor a target to clear, and gaming of targets is documented in the English health service [11]; the provisions in the Findings reduce the incentive and cannot remove it. Sixth, AICAF is not a quantitative model of optimal stopping and its rule does not weigh the cost of delay; the Copilot reading shows it returning a Hold for a product that succeeded. Where probabilities and payoffs can be estimated, an information-value analysis [42] could calibrate the floors, and I have not attempted it.

Validation designs

Controlled validation, independent raters and comparison with other methods remain future work. Table 13 names five designs and the proposition each tests; a Delphi panel and a single longitudinal case are not controlled tests, and the table says what each can show. The first design to run is use of the framework at the gate, with the tier, the levels, the challenger's rating and the date recorded before the decision, because those are the records my retrospective application lacks. The next three need no organisation to adopt the framework. The most demanding test is a comparison of AI portfolios governed by evidence gates with portfolios under conventional project review, on escalation decisions, time to production, unit economics and termination rates, with Camuffo and colleagues' 1657andomized designs as the template [13,14]. Proportionality needs its own cross-industry test, for which the EU AI Act's tiering [27] and SS1/23's risk classification [77] give the boundary conditions.

Table 13: Validation designs and the propositions they test.

Design	What is done	Tests
Use at the gate, with ratings recorded (longitudinal case study)	The framework is used on one or more initiatives through at least two gates. The tier and the component that set it, the level on each dimension, the challenger's rating, the date and the decision are recorded at each gate, before the outcome is known	Feasibility of the gates in use; whether the exposure rating, the level floor and independent challenge can be operated; P1 and P5 within the case
Retrospective coding of a historical portfolio	Two coders assign levels and tiers from the gate records of past AI initiatives and compare them with later Stop, rollback and write-off	P1 and P6; P3 and P4 where the records span G3 and G4
Structured expert panel on the rubric	A Delphi panel of practitioners and risk specialists in one sector rates the indicators in Appendix C and proposes sector benchmarks [23,73]	Content validity of the dimensions and levels; a precondition for P3 and P4
Inter-rater reliability study	Assessors rate shared evidence packs independently, with and without the rubric and the exposure rating;	P2

	agreement on levels, tiers and the binary gate result is reported with a chance-corrected coefficient for several raters and ordinal categories, such as Krippendorff's alpha [41]	
Comparative test against Stage-Gate or Agile-Stage-Gate	Matched initiatives, or the same evidence packs, are decided under each method; decisions and later reversals are compared	P5 and P6, and whether AICAF's additions change any decision

CONCLUSION

AI commercialisation is usually framed either as technology adoption or as business model innovation, and each frame is incomplete on its own; the surveys assembled here show adoption rising while the share of organisations reporting earnings impact stays flat, which I read as exposure running ahead of evidence. The AI Commercialisation Prudential Framework answers with one principle and a set of gates: commercial exposure should increase only when evidence adequacy, judged dimension by dimension against a floor set by the gate and the exposure rating, has reached the level the exposure demands and the result meets the threshold set in advance. The principle takes its approval architecture from prudential supervision, its staging from real options and validated learning, and its respect for iteration inside a pre-production stage from the Stage-Gate literature it extends; sequential decision theory gives it a lineage for the Hold and no derivation of the floors. The calibration rubric and the reading of Zillow Offers and GitHub Copilot show how the levels and gates would be applied. The retrospective application of its dimensions and outcomes to eight initiatives that I led or advised shows that past decisions can be read in those terms. None of this tests the framework. The next step is the five validation designs in Table 13, beginning with use at the gate with the ratings recorded. For an executive sponsor and a finance function, use of the framework would attach each increase in commitment to a gate record that states what was evidenced and what was assumed. For a risk function it would define a seat as challenger with a rating that stands on the record. A committee that adopts the worksheet and the evidence register will generate the data for P1 and P5 in the course of its own decisions.

Declarations

Conflict of interest: The author designed the framework described in this paper and has no commercial interest in it. The author advises organisations on product and commercial decisions, including decisions of the kind the framework addresses, and held executive or advisory roles in the organisations whose initiatives are described in the application section of the Findings. No organisation is named, and the author holds no financial interest in the organisations or products described. The insurer example is constructed, and the retrospective reading of two public cases draws only on the public documents cited. The author declares no financial interest in any source cited.

Funding: No external funding was received for this study.

Data availability: The eight initiatives in the application section are described with the permission of the organisations concerned. The organisations are unnamed, and the underlying records are confidential and are not available. Exposure tiers, evidence levels and decision dates were not recorded for those initiatives and have not been reconstructed. No other new data were created in this study. The retrospective reading of two public cases uses only the public company documents in the reference list. Every value plotted or annotated in Figures 1 to 5 is listed in Table 15, Appendix B, with its source reference number and fieldwork period; three plotted values are author renderings of verbal or graphical reports, are marked as such there, and are drawn hollow or hatched in the figures. All other material is in the reference list.

Ethical approval: Not applicable. The study used published literature, publicly available institutional sources and the author's own professional experience. That experience is the basis for describing the eight initiatives, which are described with the permission of the organisations concerned, with no organisation named and no confidential detail disclosed. The application material concerns organisational decisions, contains no personal data and identifies no individual. No human participants were recruited and no animals were involved.

Source verification: Every plotted value in Appendix B was checked against its source before use. The MIT NANDA values [15] were read from a publicly mirrored copy of the v0.1 report, because the publisher distributes the report through a gated form; the reference gives both locations, and the mirrored copy was not cross-checked against a second reproduction. The MANAGE 1.1 wording [69] was verified against the AI RMF 1.0 PDF and against the Playbook page [71], which differ in the number of "purpose". Of the works on decision theory, measurement and the lean startup, [12,40,51,56,73,79,89] were read in the original text or on the publisher's article page. Five more were read on 5 October 2026 in other copies: Hayes and Krippendorff [41] in the authors' manuscript copy, marked as submitted to and published in the journal, on the site of the Annenberg School for Communication at the University of Pennsylvania; Howard [42] in the reprint in Howard and Matheson's Readings on the Principles and Applications of Decision Analysis; Felin and colleagues [31] in the open-access copy of the article in press in Bocconi University's repository; Bevan and Hood [11] in a copy of the published article on a third-party site; and the first chapter of Dixit and Pindyck [25] in a copy on a university course page, the rest of the book not being read. Dalkey and Helmer [23], Landis and Koch [52], Lempert and colleagues [54] and Marchau and colleagues [58] are cited from the publisher's abstract or description, for what it states. Cohen [17] is cited from the bibliographic record and from the author's own summary of the paper in a 1986 Citation Classic commentary; the 1960 article was not read. The volume and page details for the two Harvard Business Review articles [12,56] are not printed on the publisher's web pages and were confirmed in library catalogue records; catalogues differ by one page on the first page of [12]. The company documents used in the retrospective case [33,35,36,37,49,66,67,76,92,93,94,95,96,97,98,99,100,101] were read on the SEC, Zillow, GitHub, arXiv and Microsoft sites on 3 October 2026. The supersession of SR 11-7 [29] and the text of its successor [30], the April 2026 update to SS1/23 [77] and the wording of SRP20 principle 3 [9] were read on the Federal Reserve, Bank of England and BIS sites on 5 October 2026. Every other source was verified against the publisher's or standard-setter's own copy.

REFERENCES

1. Adner, R., & Levinthal, D. A. (2004). What is not a real option: Considering boundaries for the application of real options to business strategy. Academy of Management Review, 29(1), 74-85. https://doi.org/10.2307/20159010
2. Amit, R., & Zott, C. (2012). Creating value through business model innovation. MIT Sloan Management Review, 53(3), 41-49. https://sloanreview.mit.edu/article/creating-value-through-business-model-innovation/ (accessed 26 September 2026)
3. Åström, J., Reim, W., & Parida, V. (2022). Value creation and value capture for AI business model innovation: A three-phase process framework. Review of Managerial Science, 16(7), 2111-2133. https://doi.org/10.1007/s11846-022-00521-z
4. Bank for International Settlements. (2024). Artificial intelligence and the economy: Implications for central banks. In BIS Annual Economic Report 2024 (Chapter III). https://www.bis.org/publ/arpdf/ar2024e3.htm (accessed 26 September 2026)
5. Bank of England & Financial Conduct Authority. (2024, November 21). Artificial intelligence in UK financial services - 2024. Bank of England. https://www.bankofengland.co.uk/report/2024/artificial-intelligence-in-uk-financial-services-2024 (accessed 26 September 2026)
6. Basel Committee on Banking Supervision. (2004). International convergence of capital measurement and capital standards: A revised framework. Bank for International Settlements. https://www.bis.org/publ/bcbs107.htm (accessed 26 September 2026)
7. Basel Committee on Banking Supervision. (2017). Basel III: Finalising post-crisis reforms. Bank for International Settlements. https://www.bis.org/bcbs/publ/d424.htm (accessed 26 September 2026)
8. Basel Committee on Banking Supervision. (2019a). The Basel Framework (consolidated framework launched 16 December 2019, continuously updated). Bank for International Settlements. https://www.bis.org/committees/bcbs/basel-framework (accessed 26 September 2026)
9. Basel Committee on Banking Supervision. (2019b). Basel Framework, SRP: Supervisory review process, chapter SRP20, Four key principles (version effective 15 December 2019). Bank for International Settlements. https://www.bis.org/basel_framework/chapter/SRP/20.htm (accessed 26 September 2026)

10. Basel Committee on Banking Supervision. (2026). Basel Framework, DIS: Disclosure requirements (version in force 1 January 2026). Bank for International Settlements. https://www.bis.org/committees/bcbs/basel-framework/standard/dis (accessed 26 September 2026)
11. Bevan, G., & Hood, C. (2006). What's measured is what matters: Targets and gaming in the English public health care system. Public Administration, 84(3), 517-538. https://doi.org/10.1111/j.1467-9299.2006.00600.x
12. Blank, S. (2013). Why the lean start-up changes everything. Harvard Business Review, 91(5), 63-72. https://hbr.org/2013/05/why-the-lean-start-up-changes-everything (accessed 3 October 2026)
13. Camuffo, A., Cordova, A., Gambardella, A., & Spina, C. (2020). A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. Management Science, 66(2), 564-586. https://doi.org/10.1287/mnsc.2018.3249
14. Camuffo, A., Gambardella, A., Messinese, D., Novelli, E., Paolucci, E., & Spina, C. (2024). A scientific approach to entrepreneurial decision-making: Large-scale replication and extension. Strategic Management Journal, 45(6), 1209-1237. https://doi.org/10.1002/smj.3580
15. Challapally, A., Pease, C., Raskar, R., & Chari, P. (2025, July). The GenAI divide: State of AI in business 2025 (Preliminary findings, v0.1). MIT NANDA, MIT Media Lab. Distributed by the publisher as a gated download from https://www.media.mit.edu/groups/nanda/overview/; values read from a publicly mirrored copy of the v0.1 PDF at https://cloudelligent.com/wp-content/uploads/2026/02/v0.1_State_of_AI_in_Business_2025_Report.pdf (both accessed 26 September 2026)
16. Chandrasekaran, A. (2026, January 26). Why 50% of GenAI projects fail: And how to beat the odds. Gartner. https://www.gartner.com/en/articles/genai-project-failure (accessed 26 September 2026)
17. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104
18. Cooper, R. G. (2008). Perspective: The Stage-Gate idea-to-launch process: Update, what's new, and NexGen systems. Journal of Product Innovation Management, 25(3), 213-232. https://doi.org/10.1111/j.1540-5885.2008.00296.x
19. Cooper, R. G. (2024). Why AI projects fail: Lessons from new product development. IEEE Engineering Management Review, 52(4), 15-21. https://doi.org/10.1109/EMR.2024.3419268
20. Cooper, R. G. (2025). Making go/kill decisions for AI in NPD: The SPARK model. IEEE Engineering Management Review, 53(5), 46-51. https://doi.org/10.1109/EMR.2024.3519998
21. Cooper, R. G., & Sommer, A. F. (2016). The Agile-Stage-Gate hybrid model: A promising new approach and a new research opportunity. Journal of Product Innovation Management, 33(5), 513-526. https://doi.org/10.1111/jpim.12314
22. Crisanto, J. C., Leuterio, C. B., Prenio, J., & Yong, J. (2024, December 12). Regulating AI in the financial sector: Recent developments and main challenges (FSI Insights on Policy Implementation No. 63). Bank for International Settlements. https://www.bis.org/fsi/publ/insights63.htm (accessed 26 September 2026)
23. Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. Management Science, 9(3), 458-467. https://doi.org/10.1287/mnsc.9.3.458
24. Deloitte. (2026, January 21). From ambition to activation: Organizations stand at the untapped edge of AI's potential, reveals Deloitte survey [Press release on the report State of AI in the enterprise 2026]. Deloitte US. https://www.deloitte.com/us/en/about/press-room/state-of-ai-report-2026.html (accessed 26 September 2026)
25. Dixit, A. K., & Pindyck, R. S. (1994). Investment under uncertainty. Princeton University Press. https://doi.org/10.1515/9781400830176
26. European Commission. (2026, August 3). AI Act [Policy page]. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed 27 September 2026)
27. European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27 (consolidated text of 27 July 2026; accessed 27 September 2026)
28. European Parliament and Council of the European Union. (2026). Regulation (EU) 2026/1744 of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on

AI). Official Journal of the European Union, L series, 2026/1744, 24 July 2026. https://eur-lex.europa.eu/eli/reg/2026/1744/oj (accessed 27 September 2026)
29. Federal Reserve Board & Office of the Comptroller of the Currency. (2011, April 4). Supervisory guidance on model risk management (SR 11-7; OCC Bulletin 2011-12; superseded on 17 April 2026 by SR 26-2). Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/boarddocs/srletters/2011/sr1107.htm (accessed 26 September 2026)
30. Federal Reserve Board, Office of the Comptroller of the Currency, & Federal Deposit Insurance Corporation. (2026, April 17). Revised guidance on model risk management (SR 26-2; OCC Bulletin 2026-13). Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm (accessed 5 October 2026)
31. Felin, T., Gambardella, A., Stern, S., & Zenger, T. (2020). Lean startup and the business model: Experimentation revisited. Long Range Planning, 53(4), Article 101889. https://doi.org/10.1016/j.lrp.2019.06.002
32. Financial Stability Board. (2024, November 14). The financial stability implications of artificial intelligence. https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/ (accessed 26 September 2026)
33. Friedman, N. (2021, June 29). Introducing GitHub Copilot: Your AI pair programmer. The GitHub Blog. https://github.blog/news-insights/product-news/introducing-github-copilot-ai-pair-programmer/ (accessed 3 October 2026)
34. Gartner. (2024, July 29). Gartner predicts 30% of generative AI projects will be abandoned after proof of concept by end of 2025 [Press release]. https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025 (accessed 26 September 2026)
35. GitHub. (2022a, June 21). GitHub Copilot is generally available to all developers. The GitHub Blog. https://github.blog/news-insights/product-news/github-copilot-is-generally-available-to-all-developers/ (accessed 3 October 2026)
36. GitHub. (2022b, December 7). GitHub Copilot is generally available for businesses. The GitHub Blog. https://github.blog/news-insights/product-news/github-copilot-is-generally-available-for-businesses/ (accessed 3 October 2026)
37. GitHub. (2023, February 14). GitHub Copilot for Business is now available. The GitHub Blog. https://github.blog/news-insights/product-news/github-copilot-for-business-is-now-available/ (accessed 3 October 2026)
38. Güner Gültekin, D., Pinarbasi, F., Yazici, M., & Adiguzel, Z. (2025). Commercialisation of artificial intelligence: A research on entrepreneurial companies with challenges and opportunities. Business Process Management Journal, 31(2), 605-630. https://doi.org/10.1108/BPMJ-10-2023-0836
39. Haakman, M., Cruz, L., Huijgens, H., & van Deursen, A. (2021). AI lifecycle models need to be revised: An exploratory study in Fintech. Empirical Software Engineering, 26(5), Article 95. https://doi.org/10.1007/s10664-021-09993-1
40. Haasnoot, M., Kwakkel, J. H., Walker, W. E., & ter Maat, J. (2013). Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. Global Environmental Change, 23(2), 485-498. https://doi.org/10.1016/j.gloenvcha.2012.12.006
41. Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1(1), 77-89. https://doi.org/10.1080/19312450709336664
42. Howard, R. A. (1966). Information value theory. IEEE Transactions on Systems Science and Cybernetics, 2(1), 22-26. https://doi.org/10.1109/TSSC.1966.300074
43. IBM Institute for Business Value. (2025, May 6). IBM study: CEOs double down on AI while navigating enterprise hurdles [Press release]. IBM Newsroom. https://newsroom.ibm.com/2025-05-06-ibm-study-ceos-double-down-on-ai-while-navigating-enterprise-hurdles (accessed 26 September 2026)
44. International Organization for Standardization & International Electrotechnical Commission. (2023). ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system (1st ed.). ISO. https://www.iso.org/standard/42001 (accessed 26 September 2026)
45. ISACA. (2026, May 5). AI use accelerates while governance and ROI lag, says new ISACA research [Press release]. https://www.isaca.org/about-us/newsroom/press-releases/2026/ai-use-accelerates-while-governance-and-roi-lag-says-new-isaca-research (accessed 26 September 2026)

46. Jobstreibizer, J., Beliaeva, T., Ferasso, M., Kraus, S., & Kallmuenzer, A. (2025). The impact of artificial intelligence on business models: A bibliometric-systematic literature review. Management Decision, 63(13), 372-396. https://doi.org/10.1108/MD-10-2024-2309
47. John, M. M., Holmström Olsson, H., & Bosch, J. (2025). An empirical guide to MLOps adoption: Framework, maturity model and taxonomy. Information and Software Technology, 183, Article 107725. https://doi.org/10.1016/j.infsof.2025.107725
48. Jorzik, P., Klein, S. P., Kanbach, D. K., & Kraus, S. (2024). AI-driven business model innovation: A systematic review and research agenda. Journal of Business Research, 182, Article 114764. https://doi.org/10.1016/j.jbusres.2024.114764
49. Kalliamvakou, E. (2022, September 7). Research: Quantifying GitHub Copilot's impact on developer productivity and happiness. The GitHub Blog. https://github.blog/news-insights/research/research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/ (accessed 3 October 2026)
50. Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., & Lahmann, A. (2024). The GenAI is out of the bottle: Generative artificial intelligence from a business model innovation perspective. Review of Managerial Science, 18(4), 1189-1220. https://doi.org/10.1007/s11846-023-00696-z
51. Knight, F. H. (1921). Risk, uncertainty and profit. Houghton Mifflin. https://archive.org/details/riskuncertaintyp00knig (accessed 3 October 2026)
52. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174. https://doi.org/10.2307/2529310
53. Lavin, A., Gilligan-Lee, C. M., Visnjic, A., Ganju, S., Newman, D., Ganguly, S., Lange, D., Baydin, A. G., Sharma, A., Gibson, A., Zheng, S., Xing, E. P., Mattmann, C., Parr, J., & Gal, Y. (2022). Technology readiness levels for machine learning systems. Nature Communications, 13(1), Article 6039. https://doi.org/10.1038/s41467-022-33128-9
54. Lempert, R. J., Popper, S. W., & Bankes, S. C. (2003). Shaping the next one hundred years: New methods for quantitative, long-term policy analysis (MR-1626-RPC). RAND Corporation. https://doi.org/10.7249/MR1626
55. Lenovo. (2026, January 27). Research from Lenovo reveals AI is paying off, yet most CIOs aren't ready for what comes next [Press release, Lenovo CIO Playbook 2026 with research by IDC]. Lenovo StoryHub. https://news.lenovo.com/pressroom/press-releases/research-reveals-ai-is-paying-off-cios-arent-ready/ (accessed 26 September 2026)
56. Magee, J. F. (1964). Decision trees for decision making. Harvard Business Review, 42(4), 126-138. https://hbr.org/1964/07/decision-trees-for-decision-making (accessed 3 October 2026)
57. Mankins, J. C. (1995, April 6). Technology readiness levels: A white paper. NASA Advanced Concepts Office, Office of Space Access and Technology. https://spacese.spacegrant.org/SEModules/Technology%20Mods/Mankins_trl.pdf (archived copy; accessed 26 September 2026)
58. Marchau, V. A. W. J., Walker, W. E., Bloemen, P. J. T. M., & Popper, S. W. (Eds.). (2019). Decision making under deep uncertainty: From theory to practice. Springer. https://doi.org/10.1007/978-3-030-05252-2
59. Martínez-Plumed, F., Gómez, E., & Hernández-Orallo, J. (2021). Futures of artificial intelligence through technology readiness levels. Telematics and Informatics, 58, Article 101525. https://doi.org/10.1016/j.tele.2020.101525
60. McGrath, R. G. (1997). A real options logic for initiating technology positioning investments. Academy of Management Review, 22(4), 974-996. https://doi.org/10.2307/259251
61. McGrath, R. G. (1999). Falling forward: Real options reasoning and entrepreneurial failure. Academy of Management Review, 24(1), 13-30. https://doi.org/10.2307/259034
62. McKinsey & Company. (2024, May 30). The state of AI in early 2024: Gen AI adoption spikes and starts to generate value. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024 (accessed 26 September 2026)
63. McKinsey & Company. (2025a, March 12). The state of AI: How organizations are rewiring to capture value. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value (accessed 26 September 2026)

64. McKinsey & Company. (2025b, December 10). AI at work but not at scale [Chart of the Week, restating the November 2025 State of AI survey]. https://www.mckinsey.com/featured-insights/charts/ai-at-work-but-not-at-scale (accessed 26 September 2026)
65. McKinsey & Company. (2026, August 25). The state of AI in 2026: On the road to ROI. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai (accessed 26 September 2026)
66. Microsoft. (2023, October 24). Microsoft fiscal year 2024 first quarter earnings conference call [Transcript]. https://www.microsoft.com/en-us/investor/events/fy-2024/earnings-fy-2024-q1 (accessed 3 October 2026)
67. Microsoft. (2024, July 30). Microsoft fiscal year 2024 fourth quarter earnings conference call [Transcript]. https://www.microsoft.com/en-us/investor/events/fy-2024/earnings-fy-2024-q4 (accessed 3 October 2026)
68. National Aeronautics and Space Administration. (2023). NASA systems engineering processes and requirements (NPR 7123.1D), Appendix E: Technology readiness levels. NASA. https://nodis3.gsfc.nasa.gov/displayDir.cfm?Internal_ID=N_PR_7123_001D_&page_name=AppendixE (accessed 26 September 2026)
69. National Institute of Standards and Technology. (2023a). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
70. National Institute of Standards and Technology. (2023b). NIST AI RMF Playbook. https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook (accessed 26 September 2026)
71. National Institute of Standards and Technology. (2023c). AI RMF Playbook: Manage. NIST Trustworthy and Responsible AI Resource Center. https://airc.nist.gov/airmf-resources/playbook/manage/ (accessed 26 September 2026)
72. National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1
73. Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. Information & Management, 42(1), 15-29. https://doi.org/10.1016/j.im.2003.11.002
74. Olechowski, A. L., Eppinger, S. D., Joglekar, N., & Tomaschek, K. (2020). Technology readiness levels: Shortcomings and improvement opportunities. Systems Engineering, 23(4), 395-408. https://doi.org/10.1002/sys.21533
75. Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), Article 114. https://doi.org/10.1145/3533378
76. Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv. https://doi.org/10.48550/arXiv.2302.06590
77. Prudential Regulation Authority. (2023). Model risk management principles for banks (Supervisory Statement SS1/23; published 17 May 2023, effective 17 May 2024; current version published 23 April 2026 following LIAF01/26, a clarifying amendment that leaves the five principles unchanged). Bank of England. https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks-ss (accessed 26 September 2026 and 5 October 2026)
78. Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. Academy of Management Review, 46(1), 192-210. https://doi.org/10.5465/amr.2018.0072
79. Ries, E. (2011). The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses. Crown Business. ISBN 978-0-307-88789-4
80. Ryseff, J., De Bruhl, B. F., & Newberry, S. J. (2024). The root causes of failure for artificial intelligence projects and how they can succeed: Avoiding the anti-patterns of AI (RR-A2680-1). RAND Corporation. https://doi.org/10.7249/RRA2680-1
81. Shepherd, D. A., & Gruber, M. (2021). The lean startup framework: Closing the academic-practitioner divide. Entrepreneurship Theory and Practice, 45(5), 967-998. https://doi.org/10.1177/1042258719899415

82. Sjödin, D., Parida, V., Palmié, M., & Wincent, J. (2021). How AI capabilities enable business model innovation: Scaling AI through co-evolutionary processes and feedback loops. Journal of Business Research, 134, 574-587. https://doi.org/10.1016/j.jbusres.2021.05.009
83. Sleesman, D. J., Conlon, D. E., McNamara, G., & Miles, J. E. (2012). Cleaning up the big muddy: A meta-analytic review of the determinants of escalation of commitment. Academy of Management Journal, 55(3), 541-562. https://doi.org/10.5465/amj.2010.0696
84. S&P Global Market Intelligence. (2025, May 30). Generative AI experiences rapid adoption, but with mixed outcomes: Highlights from VotE: AI & Machine Learning [Research highlight]. https://www.spglobal.com/market-intelligence/en/news-insights/research/ai-experiences-rapid-adoption-but-with-mixed-outcomes-highlights-from-vote-ai-machine-learning (accessed 26 September 2026)
85. Stanford Institute for Human-Centered Artificial Intelligence. (2026). AI Index Report 2026, Chapter 4: Economy. Stanford University. https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_4_economy.pdf (accessed 26 September 2026)
86. Teece, D. J. (2010). Business models, business strategy and innovation. Long Range Planning, 43(2-3), 172-194. https://doi.org/10.1016/j.lrp.2009.07.003
87. Teng, D., Ye, C., & Martinez, V. (2025). Gen-AI's effects on new value propositions in business model innovation: Evidence from information technology industry. Technovation, 143, Article 103191. https://doi.org/10.1016/j.technovation.2025.103191
88. Vallé, R., Seitz, J., Kanbach, D. K., & Schuhmacher, M. C. (2026). The use of artificial intelligence in new product development: A systematic literature review, conceptual framework, and future research agenda. Technovation, 153, Article 103541. https://doi.org/10.1016/j.technovation.2026.103541
89. Wald, A. (1945). Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics, 16(2), 117-186. https://doi.org/10.1214/aoms/1177731118
90. Weber, M., Engert, M., Schaffer, N., Weking, J., & Krcmar, H. (2023). Organizational capabilities for AI implementation: Coping with inscrutability and data dependency in AI. Information Systems Frontiers, 25(4), 1549-1569. https://doi.org/10.1007/s10796-022-10297-y
91. Wixom, B. H., Someh, I. A., & Gregory, R. W. (2021, December 16). Scaling AI to generate better and different outcomes (Research Briefing Vol. XXI, No. 12). MIT Center for Information Systems Research. https://cisr.mit.edu/publication/2021_1201_ScalingAI_WixomSomehGregory (accessed 26 September 2026)
92. Ziegler, A., Kalliamvakou, E., Simister, S., Sittampalam, G., Li, A., Rice, A., Rifkin, D., & Aftandilian, E. (2022). Productivity assessment of neural code completion. arXiv. https://doi.org/10.48550/arXiv.2205.06537
93. Zillow Group, Inc. (2018, April 12). Zillow expands Instant Offers to Phoenix; will work with agents to test buying and selling homes directly [Press release, Exhibit 99.1 to Form 8-K]. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1617640/000119312518115536/d557189dex991.htm (accessed 3 October 2026)
94. Zillow Group, Inc. (2021a, February 10). Shareholder letter, fourth quarter 2020 [Exhibit to Form 8-K]. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1617640/000161764021000007/exhibit993.htm (accessed 3 October 2026)
95. Zillow Group, Inc. (2021b, February 25). Zillow starts making cash offers for the Zestimate [Press release]. https://zillow.mediaroom.com/2021-02-25-Zillow-Starts-Making-Cash-Offers-For-the-Zestimate (accessed 3 October 2026)
96. Zillow Group, Inc. (2021c, May 4). Shareholder letter, first quarter 2021 [Exhibit to Form 8-K]. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1617640/000161764021000030/exhibit993.htm (accessed 3 October 2026)
97. Zillow Group, Inc. (2021d, August 5). Shareholder letter, second quarter 2021 [Exhibit to Form 8-K]. U.S. Securities and Exchange Commission.

https://www.sec.gov/Archives/edgar/data/1617640/000161764021000053/exhibit993.htm (accessed 3 October 2026)

98. Zillow Group, Inc. (2021e, November 2). Form 10-Q for the quarterly period ended September 30, 2021. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1617640/000161764021000087/z-20210930.htm (accessed 3 October 2026)
99. Zillow Group, Inc. (2021f, November 2). Form 8-K (Items 2.02, 2.05 and 9.01). U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1617640/000161764021000085/z-20211102.htm (accessed 3 October 2026)
100. Zillow Group, Inc. (2021g, November 2). Shareholder letter, third quarter 2021 [Exhibit to Form 8-K]. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1617640/000161764021000085/exhibit993.htm (accessed 3 October 2026)
101. Zillow Group, Inc. (2021h, November 2). Zillow Group reports third-quarter 2021 financial results & shares plan to wind down Zillow Offers operations [Press release, Exhibit 99.1 to Form 8-K]. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1617640/000161764021000085/q32021991.htm (accessed 3 October 2026)
102. Zott, C., Amit, R., & Massa, L. (2011). The business model: Recent developments and future research. Journal of Management, 37(4), 1019-1042. https://doi.org/10.1177/0149206311406265

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

APPENDIX A

AICAF assessment worksheet

The worksheet operationalises the framework for internal use; the level definitions, the two-assessor rule and the tier rating are those given in the Findings, and each entry under Key evidence refers to a line in the evidence register. Evidence is attached to each judgement and material assumptions are named.

Table 14: AICAF assessment worksheet.

Dimension	Evidence owner	Owner's level	Challenger's level	Key evidence	Material gap or assumption	Stressed at this gate?
Customer value		Not started / emerging / demonstrated / repeated	Not started / emerging / demonstrated / repeated			
Technical adequacy		Not started / emerging / demonstrated / repeated	Not started / emerging / demonstrated / repeated			
Operational adequacy		Not started / emerging / demonstrated / repeated	Not started / emerging / demonstrated / repeated			
Economic adequacy		Not started / emerging / demonstrated / repeated	Not started / emerging / demonstrated / repeated			
Risk and governance adequacy		Not started / emerging / demonstrated / repeated	Not started / emerging / demonstrated / repeated			
Organisational adequacy		Not started / emerging / demonstrated / repeated	Not started / emerging / demonstrated / repeated			
Exposure rating (Table 4)		Tier 1 / 2 / 3 on each of the four components; overall tier is the highest	Tier 1 / 2 / 3 on each component; the higher tier stands			
Gate floor (Table 6)		Minimum level on each stressed dimension for this gate and tier				
Evidence adequacy for the gate		Lowest level among stressed dimensions: at the required level / below it. Acceptance thresholds set in advance: met / not met on each stressed dimension. The gate is passed only when both hold				

Recommended decision record: Advance (with any dated evidence obligation stated), Hold, Remediate or Stop. The record identifies the next stage, the re-rated exposure, the assumptions being relied upon, the owner of unresolved gaps, the evidence required at the next review or periodic review, the date of that review, any forum override of the challenger's level with its reason, and, from Stage 3 onwards, the monitoring thresholds, the change events that would reopen a gate and the challenger's materiality determination on any change proposed.

It also carries the acceptance thresholds for the next gate, confirmed by the challenger, and any condition recorded as not applicable with the challenger's agreement.

Sensitivity prompts for the assumption record: What happens if model or inference costs rise? If adoption falls short? If error rates change after a model update? If a supplier changes its model, API or terms? If regulation tightens or a new obligation date arrives? If human-review requirements increase? If the system has to be withdrawn, what is the fallback process and who is told? Who is accountable if the system harms a customer?

APPENDIX B

Plotted values for Figures 1 to 5

Every value plotted or annotated in Figures 1 to 5, with its source reference number and the period the value describes. "Fielded" gives the fieldwork period where the source states it. Three plotted values are author renderings rather than published numbers and are marked as such: the 2026 adoption point in Figure 1 (published only as "nearly nine in ten"), the 2025 EBIT point in Figure 1 (plotted at the 2026 value on the strength of "about the same share as last year") and the general-purpose "Piloted" bar in Figure 4 Panel A (read from the ordering of labels on the source exhibit).

Table 15: Plotted and annotated values, Figures 1 to 5.

Figure	Series or item	Value	Fieldwork period (sample)	Status	Source
1	AI use in at least one business function	78%	16 to 31 Jul 2024 (n = 1,491)	Published	[63]
1	AI use in at least one business function	88%	25 Jun to 29 Jul 2025 (n = 1,993)	Published on the Chart of the Week page, which restates the November 2025 report; sample and fieldwork dates taken from the same page	[64]
1	AI use in at least one business function	89%	4 May to 8 Jun 2026 (n = 1,719)	Author rendering of "nearly nine in ten"; hollow marker	[65]
1	Any EBIT impact attributed to AI	37%	25 Jun to 29 Jul 2025 (n = 1,993)	Author rendering of "about the same share as last year"; hollow marker	[65]
1	Any EBIT impact attributed to AI	37%	4 May to 8 Jun 2026 (n = 1,719)	Published	[65]
1	AI high performers	6%	25 Jun to 29 Jul 2025 (n = 1,993)	Published as "about 6 percent"	[65]
1	AI high performers	6%	4 May to 8 Jun 2026 (n = 1,719)	Published as "remained flat at about 6 percent"	[65]
2	Global corporate AI investment, USD billion	14.57	2013	Published, Figure 4.2.1 data label	[85]
2	Global corporate AI investment, USD billion	19.04	2014	Published	[85]
2	Global corporate AI investment, USD billion	25.43	2015	Published	[85]
2	Global corporate AI investment, USD billion	33.82	2016	Published	[85]
2	Global corporate AI investment, USD billion	53.72	2017	Published	[85]
2	Global corporate AI investment, USD billion	79.62	2018	Published	[85]
2	Global corporate AI investment, USD billion	103.27	2019	Published	[85]

2	Global corporate AI investment, USD billion	221.87	2020	Published	[85]
2	Global corporate AI investment, USD billion	360.73	2021	Published	[85]
2	Global corporate AI investment, USD billion	253.25	2022	Published	[85]
2	Global corporate AI investment, USD billion	201.00	2023	Published	[85]
2	Global corporate AI investment, USD billion	253.02	2024	Published	[85]
2	Global corporate AI investment, USD billion	581.69	2025	Published; prose gives "a 129.9% increase from the previous year"; nominal or constant basis not stated in the chapter	[85]
2	Annotation: any EBIT impact attributed to AI	37%; "about the same share" in 2025	4 May to 8 Jun 2026 (n = 1,719); 25 Jun to 29 Jul 2025 (n = 1,993)	Text annotation, not a plotted series; different population from the investment series	[65]
3	Organisations abandoning the majority of AI initiatives before production, 2024 survey	17%	2024 survey (n = 1,006); comparison value printed in the 2025 report	Published (bar)	[84]
3	Organisations abandoning the majority of AI initiatives before production, 2025 survey	42%	2025 survey (n = 1,006)	Published (bar)	[84]
3	Average share of AI projects scrapped between proof of concept and broad adoption	46%	2025 survey (n = 1,006)	Text annotation only	[84]
3	Generative AI projects to be abandoned after proof of concept by end-2025	at least 30%	Prediction, 29 Jul 2024; no sample published	Prediction (hollow diamond)	[34]
3	Generative AI projects abandoned after proof of concept	at least 50%	Finding, 26 Jan 2026; no sample published	Later finding (filled diamond)	[16]
4A	Investigated: general-purpose LLMs; embedded or task-specific GenAI	80%; 60%	Jan to Jun 2025 (52 interviews, 153 survey responses, 300+ initiatives)	Published; read from a publicly mirrored copy of the v0.1 report	[15]
4A	Piloted: general-purpose LLMs; embedded or task-specific GenAI	50%; 20%	as above	General-purpose value is an author rendering read from exhibit label order (asterisk, hatched); task-specific value published	[15]
4A	Reached production: general-purpose LLMs; embedded or task-specific GenAI	40%; 5%	as above	Published ("nearly 40 percent"; "just 5 percent")	[15]
4B	Organisations that moved 40% or more of AI pilots into production	25%	Aug to Sep 2025 (n = 3,235); published Jan 2026	Published	[24]
4B	AI proofs of concept that progressed into production	46%	16 Sep to 17 Oct 2025 (n = 3,120); published Jan 2026	Published	[55]
4B	AI initiatives scaled enterprise-wide	16%	Feb to Apr 2025 (n = 2,000 CEOs); published May 2025	Published	[43]

5	Believe employees are using AI in the organisation (belief measure)	90%	2026 poll (more than 3,400 digital-trust professionals)	Published	[45]
5	Formal, comprehensive AI policy, 2026	38%	2026 poll (more than 3,400)	Published	[45]
5	Formal, comprehensive AI policy, 2025	28%	2025 poll; prior-year comparison printed in the 2026 release; 2025 sample size not given	Published	[45]
5	Mature model for agent governance	21%	Aug to Sep 2025 (n = 3,235)	Published	[24]
5	Enterprise-wide responsible-AI council or board with decision authority	18%	22 Feb to 5 Mar 2024 (n = 1,363)	Published	[62]
5	CEO responsible for overseeing AI governance	28%	16 to 31 Jul 2024 (n = 1,491); published 12 Mar 2025	Published; records where accountability sits, not a governance-lag measure	[63]

Excluded from Figure 5: the ISACA breakdown of 30 per cent "limited policy" and 25 per cent "no active policy", because it was paraphrased rather than read verbatim. Excluded from Figure 1: waves before 2024, because the high-performer definition changed (20 per cent or more of EBIT in 2020 to 2022; at least 5 per cent and "significant" in 2025 to 2026).

Milestones and legal bases for Figure 8

Table 16: Milestones drawn in Figure 8, with legal basis and marker. "As amended" means as amended by Regulation (EU) 2026/1744; all legal text read on EUR-Lex on 27 September 2026 [27,28].

Date	Milestone	Legal basis	Marker
12 July 2024	Regulation (EU) 2024/1689 published	Official Journal, L series, 12 July 2024	Circle
1 August 2024	Entry into force	Article 113, first paragraph	Circle
2 February 2025	Chapters I and II apply: scope, definitions, AI literacy and prohibited practices	Article 113(a)	Circle
2 August 2025	Chapter III Section 4, Chapter V (general-purpose AI models), Chapter VII, Chapter XII except Article 101, and Article 78 apply	Article 113(b), unchanged	Circle
24 July 2026	Regulation (EU) 2026/1744 published	Official Journal, L series, 2026/1744	Square
27 July 2026	Regulation (EU) 2026/1744 enters into force; Articles 102 to 110 apply	Article 4 of the amending regulation; Article 113(d) as amended	Square
2 August 2026	General application, including Chapter IV (Article 50 transparency obligations) and Article 101	Article 113, second paragraph, unchanged	Circle
2 December 2026	Prohibitions in Article 5(1)(ba) and (bb) apply; providers of synthetic-content systems placed on the market before 2 August 2026 comply with Article 50(2)	Article 113(a) as amended; Article 111(4) as inserted	Diamond

2 August 2027	Each Member State has at least one national AI regulatory sandbox operational (originally 2 August 2026)	Article 57(1) as amended	Diamond
2 August 2027	General-purpose AI models placed on the market before 2 August 2025 comply	Article 111(3), unchanged	Circle
2 December 2027	Chapter III Sections 1 to 3 apply to Annex III high-risk systems (originally 2 August 2026)	Article 113(c)(i) as amended	Diamond
2 August 2028	Chapter III Sections 1 to 3 apply to Annex I high-risk systems (originally 2 August 2027)	Article 113(c)(ii) as amended	Diamond
2 August 2030	High-risk systems intended for use by public authorities comply	Article 111(2); paragraph replaced by the amendment with the date unchanged	Circle
31 December 2030	AI components of the large-scale IT systems in Annex X comply	Article 111(1), unchanged	Circle

APPENDIX C

Calibration rubric for the evidence levels

Table 17 gives illustrative indicators for each dimension at three levels and one threshold example per dimension. The label on each example names the level at which the evidence must be observed before the threshold is tested; reaching the level and meeting the threshold are separate tests, and a gate needs both. Every indicator and number is illustrative: operating thresholds are set in the firm's policy, and sector benchmarks are left to the expert panel described in the Findings. The level definitions in the Findings govern where the two differ, and each condition appears under one dimension only.

Table 17: Calibration rubric: illustrative indicators by dimension and level, with a threshold example labelled by level. Illustrative only; operating thresholds are set in the firm's policy.

Dimension	Emerging	Demonstrated	Repeated	Threshold example (illustrative)
Customer value	Favourable feedback or usage from volunteers or a demonstration; no stated baseline	Outcome improvement measured once on a representative group against a stated baseline, with repeat use recorded	The same improvement in a second period or a second independent group, with no material decline	Demonstrated: handling time at least 15 per cent below the prior-quarter baseline on a representative group, with at least 60 per cent of users active each week
Technical	Task performance on synthetic or curated data, or on a single test set	Performance and error rates on a representative workload against a stated baseline, with edge-case tests and failure analysis recorded and the drift alert tested once	Results reproduced on a second independent sample or period; drift monitored over two periods with no material deterioration	Demonstrated: error rate on the representative test set no worse than the human baseline, and no severity-one failure in the edge-case suite
Operational	Integration shown in a test environment or at pilot volume only	A monitored run at target volume in the live workflow, with latency, uptime and human-review load recorded and the incident procedure exercised once	Two periods at target volume within service levels; incident procedure exercised at least twice	Demonstrated: four consecutive weeks at target volume with 99.5 per cent availability and the review queue within its service level
Economic	A cost and benefit model built on	One period of realised price or savings and of	Two periods of realised unit	Repeated: net benefit positive in two

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

	assumptions or pilot data	cost-to-serve at production volume, with the sensitivity analysis run	economics that survive the recorded sensitivities	consecutive quarters, and still positive if inference cost rises by 50 per cent
Risk and governance	Risk assessment, privacy review and supplier analysis drafted; controls specified and not yet exercised	Assessments completed and signed off by the challenger; each control (change control, access and approval controls) exercised once with the outcome recorded	Assessments refreshed in a second period with no material change; controls exercised in two periods or independent tests	Repeated: every release in two review periods passed change control, and a sample of outputs in each period showed none released without the required approval
Organisational	An owner is named; skills, support and operating model are planned	Named owner, support rota and training in operation for one period	Ownership and support in operation for two periods, including one absence or handover covered	Repeated: support requests resolved within the agreed time in two consecutive quarters, with a named owner for every dimension