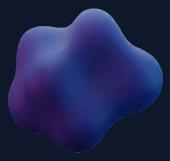

WHITEPAPER

Measuring What Matters

A universal framework for qualitative and quantitative success measurement in AI & automation products

Sam Copsey | blog.samcopsey.co.uk

March 2026 · Version 1.0

Contents

1. Executive Summary

Engineering teams building AI and automation products typically track hours logged, velocity, completion rates, and cost per item. These metrics tell you *how much* was built and *how fast*, but they tell you almost nothing about whether what was built is actually working.

> **The Measurement Problem**
>
> A 2025 MIT study found that **95% of generative AI projects fail to demonstrate ROI when measured using traditional metrics** (Challapally, Pease, Raskar & Chari, MIT Project NANDA, 2025). The projects aren't failing. The measurement is wrong.

Traditional ROI metrics measure *activity*, not *outcomes*. They miss quality, satisfaction, trust, cognitive load, and strategic value. Hard ROI from AI typically appears 12–18 months after deployment; the soft signals appear within weeks (UC Berkeley, "Squishy ROI," 2024).

Thesis: Hours saved is necessary but insufficient. Organisations that achieve sustainable ROI from AI investments use 6–10 complementary KPIs spanning multiple dimensions (Gartner, 2024). High-maturity organisations measure across 76% of their defined KPI categories, compared to 25% for low-maturity organisations.

This whitepaper introduces a **Seven Dimensions of Success** framework, synthesised from SPACE, DORA, Balanced Scorecard, Gartner, NIST, and MITRE research. It also introduces *value moments* and addresses responsible AI as a cross-cutting lens. It was built for a team delivering multiple concurrent AI products across document review, vendor research, service monitoring, and cloud management. It answers the question stakeholders actually care about: *"Is this stuff making a difference?"*

Key takeaways:

- Quantitative metrics (hours saved, velocity, cost per item) remain essential; don't discard them
- Qualitative metrics (satisfaction, trust, cognitive load) appear *before* hard ROI and predict whether it will follow
- Validated psychometric instruments exist for every dimension; you don't need to invent survey questions
- Layered measurement yields 30–50% higher returns from AI investments (McKinsey, 2024)
- Responsible AI is a cross-cutting lens across quality, trust, and strategic impact
- Measurement should centre on *value moments*: the specific interactions where value is delivered
- Implementation takes three months: baseline, instrument, report

2. The Problem with Quantitative-Only Measurement

2.1 Hours Saved Is Valuable but Insufficient

Hours saved is the most intuitive metric for automation ROI. If a process took 4 hours manually and now takes 20 minutes, you saved 3 hours and 40 minutes. Multiply by frequency, multiply by cost per hour, and you have a number for the board.

The problem is what this number misses:

- **Quality of output.** A report generated in 20 minutes that requires 2 hours of manual correction has not saved 3 hours and 40 minutes. It may have saved nothing, or created negative value through false confidence in an unchecked output.
- **User adoption.** A product that saves 100 hours per week *in theory* but is used by 2 of 50 eligible users saves only 4 hours per week *in practice*.
- **Trust dynamics.** If users don't trust AI-generated outputs, they manually verify everything, converting "time saved" into "time shifted." Anthropic's research notes that the productivity gains from AI tools depend heavily on whether users trust the output enough to reduce verification overhead (Anthropic Economic Impact Report, 2025).
- **Strategic value.** Some products don't save hours. They enable capabilities that didn't exist before. A customer research platform can provide research capability a sales team never had: hours saved = 0, value created = significant.

2.2 Cost-Per-Item Ignores Impact

A team's cost-per-item metric tells you how efficiently work items are produced. It does not tell you whether those work items produced valuable products. A team could achieve an excellent cost-per-item by delivering many low-impact items, while a team delivering fewer, higher-impact items looks worse on this metric.

2.3 Velocity Measures Activity, Not Outcomes

Consider a team with a 97.8% completion rate across six sprints, an excellent number that tells you the team delivers what it commits to. It says nothing about whether what was committed to was the right work, whether users adopted the results, or whether business outcomes improved.

> **The DORA AI Paradox**
>
> Teams using AI coding tools showed 7.5% improvement in documentation quality but 7.2% *decrease* in system stability, suggesting that speed without quality measurement creates net negative outcomes (DORA AI Impact Study, 2024).

2.4 Traditional ROI Appears Late

Hard financial ROI from AI products typically takes 12–18 months to materialise. This creates a measurement gap where leadership asks "What's the ROI?" and the honest answer is "We don't know yet" for the first year of any product's life.

During this gap, soft signals are already visible:

- **Adoption curves** show within weeks whether users find the product valuable enough to use
- **Satisfaction scores** indicate within a month whether the product meets user expectations
- **Trust levels** predict within a quarter whether users will rely on the product long-term

> **UC Berkeley: "Squishy ROI"**
>
> Products with high early adoption and satisfaction almost always deliver financial returns; products with low adoption and satisfaction almost never do, regardless of their theoretical time-saving potential (UC Berkeley Haas School of Business, 2024).

It is worth noting that this is a correlational finding from observational studies. Selection effects may contribute: products that are genuinely useful tend to achieve both high adoption and eventual financial returns, not necessarily because adoption *causes* returns. Nevertheless, the predictive relationship is strong enough to be practically useful as an early signal.

2.5 The MIT Finding: The Measurement Is Wrong, Not the Projects

When MIT researchers examined why 95% of gen AI projects failed to show ROI, they found the primary issue was **measurement methodology, not project quality**. Gen AI products often create value through:

- **Knowledge democratisation:** making expertise accessible to non-experts
- **Decision quality improvement:** better decisions, not just faster ones
- **Innovation capacity:** enabling new capabilities, not just automating old processes
- **Cognitive load reduction:** reducing mental fatigue and context-switching costs

None of these show up in a hours-saved calculation. All of them drive real business value.

Deloitte's research reinforces this: **85% of organisations achieving strong ROI from generative AI use different measurement frameworks** than they use for traditional automation (Deloitte, 2024).

3. Hours Saved: Keeping It, Improving It

Hours saved remains a core metric. The argument is not “replace hours saved with surveys.” It’s “hours saved alone gives you an incomplete and sometimes misleading picture.”

3.1 Anthropic’s Estimation Methodology

Anthropic published a rigorous methodology for estimating AI-driven productivity gains (Anthropic Economic Impact Report, 2025):

1. **Identify the task, not the job.** Break roles into discrete tasks. Estimate AI impact per task, not per role.
2. **Measure task duration before and after.** Anthropic found that AI augments approximately **57% of tasks** and fully automates approximately **43%**, with speedups varying significantly by task type, but only on the subset of tasks suitable for AI assistance.
3. **Weight by occupational time allocation.** If a task represents 20% of a role’s time and AI reduces it by 80%, the role-level impact is 16%, not 80%.
4. **Apply wage data for economic value.** Pair time savings with occupational wage data to calculate economic value.

3.2 How to Measure Hours Saved Properly

Step 1: Baseline before automation

- How many people perform this task?
- How often? (Daily, weekly, per-customer, per-document)
- How long does each instance take? (Time the actual task, not the estimate)
- What is the quality of the manual output? (Error rates, rework frequency)

Step 2: Track after automation

- Time to complete the same task using the product
- Time spent on quality validation of the automated output
- Time spent on exceptions (cases the automation can’t handle)
- Ramp-up time (the first month of usage will be slower than steady state)

Step 3: Calculate net savings

Net hours saved = (Manual time) − (Automated time + Validation time + Exception handling time + Ramp-up amortisation).

3.3 Common Pitfalls

Pitfall	Why It Matters	Mitigation
Self-reporting bias	People overestimate time saved. “This used to take me all day” often means 3 hours spread across the day.	Use time-tracking data for baselines, not retrospective estimates.
Ignoring validation time	AI outputs require human review. If a 4-hour task becomes 20-min AI + 2-hour review, savings are 1h40m, not 3h40m.	Always include validation time in post-automation measurements.
Not tracking redeployment	Saving 10 hours/week has zero value if those hours are absorbed by meetings and email.	Track what reclaimed hours are spent on.
Theoretical vs actual usage	A product that *could* save 20 hrs/week but is used for 5 hrs saves 5 hrs/week.	Multiply per-use savings by actual usage, not potential usage.
Ignoring quality differences	Automated output may be higher or lower quality than manual.	Measure quality before and after. Include rework time.

3.4 Enhancement: Hours Redeployed to Strategic Work

Category	Description	Value Multiplier
Strategic work	New product development, innovation, capability building	High. Directly generates future value
Quality improvement	Better testing, documentation, architecture, security	Medium. Reduces future costs and risks
Absorbed by BAU	Meetings, email, admin, unplanned support	Low. Hours saved but value not captured

4. Value Moments: Where Products Deliver Impact

4.1 What Is a Value Moment?

A **value moment** is the specific user interaction where a product delivers its intended value. It is the point at which time is saved, a decision is improved, a risk is detected, or a capability is unlocked.

Value moments matter because they focus measurement on what actually counts. A product can have high login rates, strong MAU numbers, and respectable session durations while delivering very little value, if users are logging in out of habit, checking dashboards without acting on them, or generating outputs they don't trust enough to use. Conversely, a product with low overall usage volume might deliver enormous value at a single, high-stakes interaction point.

Examples:

- For an **AI document reviewer**: the value moment is the point at which a user receives a reviewed document and accepts the output without significant manual correction. The value is not "document submitted" or "review started." It is "review completed and accepted."
- For a **monitoring system**: the value moment is the detection of an incident that would otherwise have been missed or detected later. The value is not "dashboard viewed." It is "incident detected before SLA breach."
- For a **research platform**: the value moment is when a user acts on a research output, using it to inform a decision, a proposal, or a customer conversation. The value is not "report generated." It is "research influenced a decision."
- For a **ticket automation system**: the value moment is the successful creation of a correctly categorised, correctly routed ticket without human intervention. The value is "ticket created correctly," not "automation triggered."

4.2 Why Value Moments Matter for Measurement

Traditional usage metrics (MAU, session count, feature utilisation) measure *engagement*. Value moments measure *impact*. The distinction matters because:

1. **Hours saved should be calculated per value moment, not per login.** If a product saves 20 minutes per document review, multiply by the number of reviews completed and accepted (value moments), not the number of sessions.
2. **Trust and satisfaction surveys should be administered at or near value moments.** A CES question asked after a user receives and acts on an AI output is far more informative than one asked at a random time.
3. **Adoption metrics should track value moment frequency, not just access.** A user who generates 3 research reports per week is demonstrating more value than one who logs in daily but generates nothing.
4. **Quality metrics should be anchored to value moments.** AI Output Accuracy (acceptance rate) is meaningful only when measured at the value moment: was the output accepted and used, or corrected and used, or rejected?

4.3 Identifying Value Moments for Your Products

For each product, answer three questions:

1. **What is the single interaction where value is delivered?** Strip away setup, navigation, and configuration. What is the core moment?
2. **How can you observe this moment in telemetry?** What event, API call, or user action marks the value moment?
3. **What is the counterfactual?** What would the user have done without the product? How long would it have taken? What quality would it have achieved?

Product Archetype	Value Moment	Observable Event	Counterfactual
AI Document Reviewer	Review completed and accepted	Accept/submit event after AI review	Manual review: 2-4 hours per document
Research Platform	Research output used in a decision	Report downloaded or shared externally	Manual research: days to weeks
Knowledge Agent	Question answered satisfactorily	User did not escalate or re-ask	Ask a colleague, search documentation: 15-30 minutes
Monitoring System	Incident detected proactively	Alert fired before customer report or SLA breach	Manual detection: hours to days
Ticket Automation	Ticket created correctly without intervention	Auto-created ticket with no manual edit within 24 hours	Manual ticket creation: 5-15 minutes per ticket
Customer Reporting	Report delivered to customer	Report sent/published event	Manual report compilation: 2-8 hours

4.4 Integrating Value Moments into the Framework

Throughout the rest of this framework, metrics and instruments should be anchored to value moments wherever possible:

- **Dimension 2 (Quality):** Accuracy measured at value moments, not in isolation
- **Dimension 3 (Adoption):** Tracked as value moment frequency, not login count
- **Dimension 4 (Satisfaction):** Survey triggers placed at or near value moments
- **Dimension 5 (Trust):** Trust calibrated against value moment outcomes
- **Dimension 7 (Strategic Impact):** Hours saved calculated per value moment

5. Framework: Seven Dimensions of Success

The following framework synthesises research from SPACE (Microsoft Research), DORA (Google), Balanced Scorecard (Kaplan & Norton), Gartner's AI Value Metrics, NIST's Trustworthiness Framework, and MITRE's AI Maturity Model into seven measurable dimensions.

Each dimension captures a different facet of product and team success. Together, they provide a layered view that shows *what* was built (delivery), *how well* it works (quality), *whether people use it* (adoption), *how they feel about it* (satisfaction, trust), *what it costs them cognitively* (cognitive load), and *what strategic value it creates* (business impact).

#	Dimension	Type	What It Answers	Key Frameworks
1	**Delivery Performance**	Quantitative	Are we shipping reliably and predictably?	DORA, SPACE (Activity)
2	**Product Quality & Reliability**	Quant + Qual	Does what we ship work correctly and consistently?	DORA, NIST, Balanced Scorecard
3	**User Adoption & Engagement**	Quant + Qual	Are people actually using what we built?	Gartner, Balanced Scorecard, Forrester
4	**User Satisfaction & Experience**	Qualitative	Do users find the product useful and usable?	SPACE, SUS, NPS, CES
5	**Trust & Confidence**	Qualitative	Do users trust the AI outputs enough to act on them?	NIST, S-TIAS, MITRE
6	**Cognitive Load & Dev Experience**	Qualitative	Does the product reduce or increase mental burden?	NASA-TLX, SPACE, GitHub Studies
7	**Strategic & Business Impact**	Quant + Qual	Is this driving outcomes that matter to the business?	Balanced Scorecard, Gartner, Forrester

Why Seven?

Gartner's research found that organisations achieving sustainable AI ROI use 6–10 complementary KPIs. Fewer than 6 leaves blind spots; more than 10 creates measurement overhead. Seven dimensions with 2–4 KPIs each yields approximately 20 total metrics, enough for a complete picture without drowning in data.

6. Per-Dimension Deep Dive

6.1 Delivery Performance

What it measures: The team's ability to ship work reliably, predictably, and at a sustainable pace.

Why it matters: Delivery performance is the foundation. If the team can't ship consistently, nothing else matters; there's nothing to measure satisfaction or adoption *of*.

KPI	Definition	How to Measure	Target	Cadence
Sprint Completion Rate	% of planned parent items completed per sprint	Sprint data: Completed / Planned	> 90%	Per sprint
Velocity (Completed Items)	Number of parent items completed per sprint	Sprint data	Stable or increasing trend	Per sprint
Velocity Delta	Completed − Planned items per sprint	Sprint data	0 (target), positive = over-delivery	Per sprint
Cycle Time	Median days from item creation to closure	Work item dates	Decreasing trend	Per sprint
Deployment Frequency	How often changes reach production	Release/deployment logs	Weekly or more frequent	Monthly
Lead Time for Changes	Time from code commit to production deployment	Git commit → deployment date	< 1 week	Monthly

Score	Sprint Completion	Velocity Delta	Cycle Time Trend
Green	≥ 90%	−2 to +5	Stable or decreasing
Amber	75–89%	−5 to −3	Slight increase
Red	< 75%	< −5	Sustained increase

Benchmarks: DORA Elite performers: deployment frequency = on-demand, lead time = less than one hour, change failure rate = 0–15%. For internal product teams, the relevant comparison is within the team over time.

6.2 Product Quality & Reliability

What it measures: Whether the products work correctly, produce reliable outputs, and minimise errors and rework.

Why it matters: Quality determines whether hours saved are *real* savings or *shifted* costs. For AI products specifically, output quality is the single biggest driver of user trust and sustained adoption.

KPI	Definition	How to Measure	Target	Cadence
Defect Rate	Bugs filed per sprint per product	Bug count, filtered by product	< 2 per product per sprint	Per sprint
Change Failure Rate	% of deployments causing incidents	Incident logs, hotfix count	< 15% (DORA Elite)	Monthly
AI Output Accuracy	% of AI outputs accepted without correction	User survey + product telemetry	> 80% acceptance rate	Monthly
Rework Rate	% of items reopened or requiring follow-up	Items closed then reopened	< 10%	Per sprint
Incident Resolution Time	Mean time to resolve production incidents	Incident ticket duration	< 4h (P1), < 24h (P2)	Monthly
Error Consistency	Variance in output quality across similar inputs	Spot-check 10 random outputs/month	Low variance (σ < 1)	Monthly

Score	Defect Rate	AI Output Accuracy	Rework Rate
Green	0–1 bugs/sprint	≥ 85%	< 5%
Amber	2–3 bugs/sprint	70–84%	5–15%
Red	≥ 4 bugs/sprint	< 70%	> 15%

Per-product applicability: All products, with AI Output Accuracy particularly critical for AI document reviewers (quality), research tools (accuracy), knowledge agents (correctness), and transcription tools (fidelity).

Data sources:

- Bug items filtered by product tag
- Product telemetry: accept/edit/reject rates on AI-generated outputs
- Monthly spot-check evaluations (10 random outputs per product)
- User-reported corrections via feedback channels

> **Important Limitation: Spot-Check Validation**
>
> Acceptance without correction does not necessarily mean the output was accurate. Users may accept outputs without thorough verification, particularly as trust increases. To mitigate this, pair acceptance tracking with a **spot-check validation layer**: independently verify a random sample of 10 accepted outputs per product per month. Compare spot-check accuracy against the acceptance rate to detect divergence.

6.3 User Adoption & Engagement

What it measures: Whether the intended users are actually using the product, how frequently, and how deeply.

Why it matters: Adoption is the bridge between theoretical ROI and actual ROI. A product in production with zero active users has zero value.

KPI	Definition	How to Measure	Target	Cadence
Monthly Active Users (MAU)	Unique users who interact per month	Product telemetry / auth logs	> 50% of eligible users	Monthly
Adoption Rate	MAU / Total eligible users	MAU ÷ eligible user count	> 60% within 3 months	Monthly
Time to Value (TTV)	Days from first use to regular use	User activity logs	< 14 days	Per user cohort
Feature Utilisation	% of core features used by average user	Product telemetry	> 70% of core features	Quarterly
Retention Rate	% of users still active at 30/60/90 days	Cohort analysis	> 80% at 30 days	Monthly
Usage Volume	Total units processed per month	Product telemetry	Growing trend	Monthly

Score	Adoption Rate (Launch, 0-3 months)	Adoption Rate (Steady State, 6+ months)	TTV	30-Day Retention
Green	≥ 30%	≥ 60%	≤ 7 days	≥ 80%
Amber	15–29%	30–59%	8–21 days	60–79%
Red	< 15%	< 30%	> 21 days	< 60%

> **Benchmark Context**
>
> Industry benchmarks for internal enterprise tool adoption are typically 20-30% in the first quarter. The targets above reflect this reality: launch-phase targets are deliberately lower than steady-state expectations. A product achieving 25% adoption in month 2 is on track, not underperforming.

Mapping usage metrics to product types

Product Type	Primary Usage Metric	Eligible User Base
AI Document Reviewer	Documents reviewed per month	Pre-sales, engineering
Research Platform	Research reports generated	Sales team
Knowledge Agent	Queries answered per week	All internal staff
Customer Reporting Tool	Reports generated per month	Service delivery team, customers
Ticket Automation	Tickets auto-created per week	Helpdesk team
Monitoring System	Incidents detected per week	Service desk, operations
Transcription Tool	Calls transcribed per month	Pre-sales, solutions architects
Analytics/Reporting	Summaries generated per month	Finance, management

6.4 User Satisfaction & Experience

What it measures: How users feel about the product: whether they find it useful, usable, and worth recommending.

Why it matters: Satisfaction is the leading indicator of adoption sustainability.

KPI	Definition	How to Measure	Target	Cadence
Net Promoter Score (eNPS)	"How likely to recommend to a colleague?" (0–10)	Single-question survey	> 20 (Good), > 50 (Excellent)	Quarterly
System Usability Scale (SUS)	10-question standardised assessment	SUS questionnaire	> 68 (average), > 80 (good)	Quarterly
Customer Effort Score (CES)	"How easy was it to accomplish your task?" (1–7)	Post-interaction survey	> 5.5	Monthly
Overall Satisfaction	"How satisfied are you with [product]?" (1–5)	Single-question survey	> 4.0	Monthly
Qualitative Feedback Themes	Top 3 positive/negative themes	Open-ended + thematic analysis	Positive outweigh negative	Quarterly

Score	eNPS	SUS	CES
Green	> 30	> 75	> 5.5
Amber	0–30	50–75	4.0–5.5
Red	< 0	< 50	< 4.0

6.5 Trust & Confidence

What it measures: Whether users trust AI-generated outputs enough to act on them without excessive manual verification.

Why it matters: Trust is the unique challenge of AI products. If users don't trust the outputs, they revert to manual processes, and the time savings evaporate.

KPI	Definition	How to Measure	Target	Cadence
S-TIAS Trust Score	3-item validated trust in automation scale	S-TIAS questionnaire (7-point Likert)	> 5.0	Quarterly
Decision Confidence	"How confident in the accuracy of this output?" (1–7)	Post-interaction micro-survey	> 5.0	Monthly
Override Rate	% of AI outputs manually overridden	Product telemetry	< 20%	Monthly
Verification Behaviour	Time spent manually checking AI outputs	Time-tracking or self-report	Decreasing trend	Quarterly
Transparency Satisfaction	"I understand why the product produced this output" (1–7)	Survey question	> 5.0	Quarterly

Score	S-TIAS Score	Override Rate	Decision Confidence
Green	> 5.5	< 15%	> 5.5
Amber	4.0–5.5	15–30%	4.0–5.5
Red	< 4.0	> 30%	< 4.0

> **DORA's Trust Calibration Insight**
>
> The target is not maximum trust; it is **appropriate trust**, calibrated to the actual reliability of the output. Users who trust AI outputs *too much* skip necessary verification; users who trust *too little* negate the time savings.

6.6 Cognitive Load & Developer Experience

What it measures: The mental effort required to use products, complete work, and maintain focus.

Why it matters: Each context switch costs **approximately 23 minutes** of recovery time (Mark et al., 2008). For a team managing multiple concurrent products, reducing unnecessary cognitive load may be worth more than any individual time saving.

> **GitHub Copilot Studies**
>
> Developers using AI assistance reported being **60–75% more fulfilled** and **73% reported staying focused** during complex tasks. Cognitive load reduction, not just time savings, drives perceived value (GitHub, 2023).

KPI	Definition	How to Measure	Target	Cadence
NASA-TLX Composite	6-dimension cognitive load assessment	NASA-TLX questionnaire (0–100)	< 50 composite	Quarterly
Context Switches per Day	Times switching between unrelated tasks	Self-report or calendar analysis	< 8/day	Monthly
Flow State Frequency	"How often 2+ hrs uninterrupted focus?"	Weekly self-report (1–5)	> 3 (often)	Monthly
Tool Overhead	Time on tooling vs productive work	Time-tracking or self-report	< 15%	Quarterly
Developer Satisfaction (SPACE)	11-item survey across 5 dimensions	SPACE questionnaire	> 3.5 per dimension	Quarterly
Frustration Level	"How frustrated with current tools?" (1–7)	Single-question survey	< 3.0	Monthly

Score	NASA-TLX Composite	Context Switches	Flow State Frequency
Green	< 40	< 6/day	> 3.5
Amber	40–60	6–10/day	2.5–3.5
Red	> 60	> 10/day	< 2.5

6.7 Strategic & Business Impact

What it measures: Whether the products are driving outcomes that matter to the business: revenue, capability, competitive advantage.

Why it matters: This is the "so what?" dimension. A team can deliver reliably and score well on satisfaction, but if the products don't contribute to business outcomes, the investment isn't justified.

KPI	Definition	How to Measure	Target	Cadence
Hours Saved (Validated)	Net hours saved per product, validated by users	User surveys + telemetry	Growing trend	Monthly
Hours Redeployed Strategically	% of reclaimed hours used for strategic vs BAU work	Time-tracking categorisation	> 50% strategic	Quarterly
Revenue Attribution	Revenue enabled or influenced by products	CRM flags, retention data	Growing contribution	Quarterly
Innovation Capacity	% of engineering time on new capabilities	Sprint item categorisation	> 50% new capabilities	Per sprint
Knowledge Democratisation	Non-experts performing expert-level tasks	Usage logs + surveys	Growing user base	Quarterly
Time-to-Decision	Average time from decision initiation to action	Process measurement	Decreasing trend	Quarterly
Cost Avoidance	Incidents prevented, SLA breaches avoided	Incident data comparison	Growing avoidance value	Quarterly

Score	Hours Saved (Validated)	Innovation Capacity	Revenue Attribution
Green	Growing, > 10 hrs/week/product	> 60%	Measurable contribution
Amber	Stable, 5–10 hrs/week/product	40–60%	Indirect or early-stage
Red	Declining or < 5 hrs/week/product	< 40%	No measurable contribution

Mapping strategic value by product archetype

Product Archetype	Primary Strategic Value	Revenue Attribution Path
AI Document Reviewer	Quality improvement (better proposals)	Win rate on reviewed proposals
Research Platform	Revenue enablement (sales research)	Deals flagged as “product-assisted”
AI Infrastructure Layer	Enables other AI products (platform value)	Indirect. Unlocks downstream products
Knowledge Agent	Knowledge democratisation	Onboarding speed, pre-sales efficiency
Customer Reporting Tool	Customer retention (proactive reporting)	Retention rates, expansion revenue
Ticket Automation	Operational efficiency	Capacity freed for higher-value work
Monitoring System	Risk reduction (faster incident detection)	SLA breach prevention
Transcription/Analysis Tool	Decision quality (scoping accuracy)	Scoping accuracy → profitability

Cost of Value Delivery

While not a scored KPI in this framework, teams should monitor the operational cost of each product: API consumption, compute resources, licensing, and support overhead. A product can score green across all seven dimensions while consuming more in API costs than the value it delivers. Track cost-per-value-moment (total monthly operating cost divided by value moment count) and review quarterly. If cost-per-value-moment exceeds the value of the manual alternative, the product's economics need revisiting regardless of how well it scores on other dimensions.

7. Responsible AI: A Cross-Cutting Lens

Why This Is Not an Eighth Dimension

Responsible AI is not a separate measurement dimension because it is not independent of the other seven. It is a lens through which several existing dimensions should be examined. Bias detection is a quality concern. Explainability is a trust concern. Data governance is a strategic and regulatory concern. Treating responsible AI as a standalone metric risks creating a checkbox exercise disconnected from the product's actual measurement programme.

Instead, this section provides a cross-cutting assessment that maps responsible AI considerations to the dimensions where they are measured.

The Regulatory Context

> **Why This Matters**
>
> For organisations in regulated sectors (healthcare, government, defence, education), responsible AI is not optional. Procurement frameworks increasingly require evidence of bias assessment, explainability, data governance, human oversight, and incident response. The EU AI Act (effective 2025-2026) and the UK's pro-innovation AI regulatory framework both emphasise these principles. Even for internal tools, demonstrating responsible AI practices strengthens the case for continued investment and reduces organisational risk.

Key requirements appearing in procurement frameworks:

- **Bias assessment:** Has the AI been tested for demographic, contextual, or systematic bias in its outputs?
- **Explainability:** Can the AI explain (or can its operators explain) why it produced a given output?
- **Data governance:** Is the data used by the AI stored, processed, and retained in accordance with data protection requirements?
- **Human oversight:** Is there a human in the loop for consequential decisions?
- **Incident response:** Is there a process for identifying and responding to AI failures, hallucinations, or harmful outputs?

Mapping Responsible AI to the Seven Dimensions

Responsible AI Concern	Relevant Dimension	What to Measure	How
Output bias	Dimension 2 (Quality)	Are outputs consistently accurate across different input types, users, or contexts?	Spot-check outputs across varied inputs. Flag systematic patterns in errors.
Explainability	Dimension 5 (Trust)	Can users understand why the AI produced a given output?	Transparency Satisfaction survey item (already in framework). Review override patterns for unexplained outputs.
Data governance	Dimension 7 (Strategic Impact)	Is data handled in compliance with organisational and regulatory requirements?	Quarterly data governance review. Document data flows per product.
Human oversight	Dimension 2 (Quality)	Are high-stakes outputs reviewed by a human before action?	Track % of outputs in high-stakes categories that receive human review.
Fairness	Dimension 3 (Adoption)	Is the product equally accessible and useful to all intended user groups?	Disaggregate adoption and satisfaction data by user role, department, or demographic where possible.
Incident response	Dimension 2 (Quality)	Is there a documented process for responding to AI failures?	Maintain an AI incident log. Review quarterly.

Responsible AI Assessment Checklist

Administer this checklist per product, quarterly. Each item is scored Yes / Partial / No.

1. **Bias testing:** Have we tested this product's outputs for systematic errors across different input types in the last quarter?
2. **Explainability:** Can a non-technical user understand why this product produced a given output (either through the product's UI or through documentation)?
3. **Data compliance:** Is the data this product processes stored and retained in accordance with our data governance policies?
4. **Human oversight:** For outputs that inform consequential decisions, is human review required before action?
5. **Incident log:** Have we documented any AI failures, hallucinations, or unexpected outputs this quarter, and have they been resolved?
6. **User consent:** Do users understand that they are interacting with an AI system and what data is being processed?

Scoring: Products with 6/6 Yes are operating responsibly. Products with any "No" items require remediation planning. Products with "Partial" items should have improvement targets for the next quarter.

This checklist is deliberately lightweight. For organisations requiring deeper responsible AI assessment, refer to the NIST AI Risk Management Framework (AI 100-1, 2023) and the MITRE AI Maturity Model's Ethics and Responsible AI pillar.

8. Qualitative Measurement Instruments

This section provides the exact instruments, questions, scoring methodologies, and benchmarks. All instruments are validated. Use them directly without modification.

8.1 Net Promoter Score / Employee Net Promoter Score (eNPS)

> *"On a scale of 0 to 10, how likely are you to recommend [product name] to a colleague?"*

Scoring: Promoters (9–10), Passives (7–8), Detractors (0–6). eNPS = % Promoters − % Detractors.

Score Range	Interpretation
> 50	Excellent. Strong advocacy
20–50	Good. More promoters than detractors
0–20	Acceptable. Roughly balanced
< 0	Concerning. More detractors than promoters

> *"What is the primary reason for your score?" (Follow-up, required for actionability)*

8.2 System Usability Scale (SUS)

10-question assessment validated across 500+ studies (Brooke, 2013). Score: 0–100. Key benchmark: 68 = 50th percentile (average).

Scoring: Odd items: response − 1. Even items: 5 − response. Sum all × 2.5.

SUS Score	Percentile	Grade	Interpretation
> 80	90th+	A	Excellent
68–80	50th–89th	B–C	Good to average
50–67	15th–49th	D	Below average
< 50	Below 15th	F	Poor

8.3 Customer Effort Score (CES)

> *"[Product name] made it easy to accomplish my task." Scale: 1–7.*

Score Range	Interpretation
> 6.0	Excellent. Very low effort
5.0–6.0	Good. Acceptably easy
4.0–4.9	Concerning. Noticeable friction
< 4.0	Poor. The product is adding effort

8.4 Short Trust in Automation Scale (S-TIAS)

3-item validated scale for AI/automation trust, the short form of Körber's TiA questionnaire (McGrath, Lack, Tisch & Duenser, *Frontiers in Artificial Intelligence*, 8, 2025. DOI: 10.3389/frai.2025.1556737). 7-point Likert. Average the responses.

Score Range	Interpretation
> 5.5	High trust. Users rely on outputs
4.0–5.5	Moderate trust. Selective verification
2.5–3.9	Low trust. Verify everything (negates savings)
< 2.5	Very low trust. Users avoid the product

8.5 Decision Quality Assessment

Five metrics from the Glean AI Decision Quality Framework (2024). Each scored 1–5.

Metric	Question
Accuracy	"The output was factually correct and free of errors"
Relevance	"The output addressed exactly what I needed"
Coherence	"The output was well-structured and logically organised"
Helpfulness	"The output saved me time and effort"
User Confidence	"I feel confident acting on this output without further verification"

8.6 NASA Task Load Index (NASA-TLX)

6-dimension cognitive load measure. Each dimension rated 0–100. Raw TLX = average of all six.

Composite Score	Interpretation
< 30	Low. Sustainable pace
30–50	Moderate. Within healthy range
50–70	High. Risk of fatigue and errors
> 70	Very high. Unsustainable

8.7 SPACE Developer Experience Survey

11 items across 5 dimensions (Forsgren et al., ACM Queue, 2021). 5-point Likert scale. Target > 3.5 per dimension.

8.8 AI Maturity Self-Assessment

Based on the MITRE AI Maturity Model (2024). Six pillars, five maturity levels.

Pillar	What to Assess
Ethical, Equitable and Responsible Use	Assess for bias? Users can challenge AI decisions?
Strategy and Resources	Clear AI strategy? Investment decisions governed?
Organisation	AI skills? Organisation open to AI-driven change?
Technology Enablers	Reliable infrastructure for AI workloads?
Data	Data clean, documented, accessible?
Performance and Application	Can you prove your AI products are working?

Level	Name	Description
1	**Initial**	Ad hoc. No systematic approach.
2	**Developing**	Some structure. Pilot projects. Basic policies.
3	**Defined**	Documented processes. Consistent approach.
4	**Managed**	Metrics-driven. Continuous improvement.
5	**Optimising**	Industry-leading. AI embedded in strategy.

Cadence: Every 6 months. Use the blank template below to score your own organisation.

9. Measurement at Small Scale

The Small-n Challenge

The validated instruments in this framework (SUS, eNPS, S-TIAS, NASA-TLX) were developed and validated in studies with sample sizes typically ranging from 12 to several hundred participants. When applied to small teams (5-10 engineers) or small user bases (2-15 users per product), the statistical properties that make these instruments reliable at scale do not hold.

Specific risks at small sample sizes:

- **eNPS** requires approximately 30 respondents for statistical stability. With 5 respondents, a single detractor can swing the score by 40 points.
- **SUS** is considered reliable at $n \geq 8$ (Tullis & Stetson, 2004), but confidence intervals widen dramatically below $n = 12$.
- **S-TIAS** (3 items) is more robust at small n than multi-item instruments, but should still be interpreted cautiously below $n = 10$.

How to Interpret Small-Sample Scores

1. **Treat scores as directional signals, not precise measurements.** An eNPS of +20 vs +40 is meaningless with 5 respondents. An eNPS that is consistently positive vs consistently negative over three quarters is meaningful.
2. **Track longitudinal trends over point-in-time snapshots.** A single SUS score of 72 tells you little. Three consecutive quarters of 72, 68, 61 tells you usability is declining. The trend matters more than any individual number.
3. **Track individual S-TIAS trajectories (anonymised).** Rather than aggregating 5 trust responses into one number, track each respondent's S-TIAS score over time (anonymised by consistent ID, not by name). One person's trust score dropping from 6.0 to 3.5 is a stronger signal than a group average moving from 5.2 to 4.8.
4. **Supplement eNPS scores with qualitative follow-ups.** At small scale, a 15-minute conversation with a detractor reveals more than the eNPS number itself. Use the follow-up question ("What is the primary reason for your score?") as a starting point for structured interviews.

5. **Report full CES distributions, not just the average.** With 5 CES responses, report all 5 scores alongside the average. “Scores: 4, 5, 5, 6, 2” is more informative than “Average: 4.4” because it reveals that one respondent found the product significantly harder to use.

Minimum Viable Measurement

For teams and user bases below 10 people, consider a simplified measurement approach:

Instead of...	Use...	Why
Full SUS (10 questions)	3-question usability check: “easy to use,” “gets the job done,” “would recommend”	Reduces burden, still captures signal
Quarterly eNPS per product	Quarterly “traffic light” retrospective per product (team votes Green/Amber/Red)	More meaningful at n < 10 than a numeric score
Formal NASA-TLX (6 dimensions)	Single question: “How manageable is your workload this sprint?” (1-5)	Captures the key signal without the overhead
Per-product satisfaction surveys	Rotating deep-dive: one product per quarter gets a thorough review, others get a single question	Reduces fatigue dramatically

The goal is not to compromise on measurement but to choose instruments that produce reliable signals at the sample sizes available.

10. Per-Product Measurement Plan

This section demonstrates how the seven dimensions map to different product archetypes. Rather than prescribing a single plan, it provides three worked examples that illustrate how to prioritise dimensions and select KPIs based on product type.

Measurement Priority by Product Archetype

The priority of each dimension varies by product type. Use this matrix as a starting point, then adjust based on your product's maturity, user base, and strategic importance.

Product Archetype	Delivery	Quality	Adoption	Satisfaction	Trust	Cognitive Load	Strategic Impact
AI Content Generator	Med	**High**	**High**	**High**	**High**	Med	Med
AI Research Platform	Med	**High**	**High**	**High**	**High**	Low	**High**
AI Infrastructure Layer	**High**	**High**	Low	Low	Low	Low	**High**
Knowledge Agent	Med	**High**	**High**	**High**	**High**	Med	Med
Customer Reporting Tool	**High**	**High**	**High**	**High**	Med	Low	**High**
Ticket Automation	Med	Med	Med	Med	Low	Low	Med
Monitoring & Alerting System	**High**	**High**	Med	Med	Med	Low	**High**
Transcription/Analysis Tool	Med	**High**	**High**	**High**	**High**	Med	Med
Analytics/Reporting	Low	Med	Med	Med	Low	Low	Low

High = primary focus, Medium = monitor regularly, Low = measure at lower cadence

Worked Example 1: AI Content Generator

An AI product that reviews, generates, or transforms documents (proposals, reports, contracts). Trust and quality are paramount because users must be confident the output is accurate before submitting it externally.

Priority Dimensions

Dimension	Priority	Rationale
Quality	**High**	Output errors in external-facing documents carry reputational risk
Trust	**High**	Users must trust output enough to submit without full manual re-review
Adoption	**High**	Value is zero if eligible users do not adopt the tool
Satisfaction	**High**	Leading indicator of sustained adoption
Strategic Impact	Med	Hours saved and quality improvement both contribute
Delivery	Med	Standard sprint tracking sufficient
Cognitive Load	Med	Relevant if tool creates new review burden

Recommended KPIs

KPI	Source	Cadence	Owner
Documents processed per month	Product telemetry	Monthly	Product owner
AI output accuracy (acceptance rate)	Accept/edit/reject tracking	Monthly	Product owner
SUS score	User survey	Quarterly	Engineering Manager
S-TIAS trust score	User survey	Quarterly	Engineering Manager
Decision quality (5 metrics)	Spot-check evaluation	Monthly	Product owner
Hours saved vs manual review	User survey + baseline	Monthly	Engineering Manager
eNPS	User survey	Quarterly	Engineering Manager

Who to survey: All users who submit content for AI review or generation.

Worked Example 2: Monitoring & Alerting System

A system that monitors infrastructure, services, or customer environments and detects incidents proactively. Delivery reliability and quality (false positive rates) are the primary concerns.

Priority Dimensions

Dimension	Priority	Rationale
Delivery	**High**	System must be always-on; downtime means missed incidents
Quality	**High**	False positives erode trust; false negatives mean missed incidents
Strategic Impact	**High**	Directly prevents SLA breaches and customer impact
Trust	Med	Operations teams must trust alerts enough to act immediately
Adoption	Med	Usage is semi-automatic; focus on coverage, not login counts
Satisfaction	Med	Operations team experience matters for sustained use
Cognitive Load	Low	Minimal direct interaction; alerts should reduce cognitive burden

Recommended KPIs

KPI	Source	Cadence	Owner
Incidents detected per week	Product telemetry	Weekly	Product owner
Mean time to detect (MTTD)	Incident timeline data	Monthly	Product owner
False positive rate	Incident review	Monthly	Product owner
SLA breaches prevented	Before/after comparison	Quarterly	Engineering Manager
Operations team satisfaction	Team survey	Quarterly	Engineering Manager
Customer satisfaction (incident communication speed)	Customer survey	Quarterly	Account management

Who to survey: Operations/service desk team, managed services team, customers.

Worked Example 3: AI Research Platform

A platform that generates research reports, analyses vendors, or compiles market intelligence using AI. Trust and strategic impact are critical because research outputs directly inform business decisions.

Priority Dimensions

Dimension	Priority	Rationale
Quality	**High**	Research accuracy directly affects decision quality
Trust	**High**	Users must trust research enough to act on it
Strategic Impact	**High**	Research enables revenue and strategic decisions
Adoption	**High**	Value depends on the sales/strategy team using it regularly
Satisfaction	**High**	Leading indicator of adoption sustainability
Delivery	Med	Standard sprint tracking sufficient
Cognitive Load	Low	Research tool should be straightforward to use

Recommended KPIs

KPI	Source	Cadence	Owner
Research reports generated per month	Product telemetry	Monthly	Product owner
Report accuracy (spot-check)	Manual evaluation	Monthly	Product owner
SUS score	User survey	Quarterly	Engineering Manager
eNPS	User survey	Quarterly	Engineering Manager
CES (post-report)	Micro-survey	Per use	Product owner
Revenue attribution (research-assisted deals)	CRM flag	Quarterly	Sales team
S-TIAS trust score	User survey	Quarterly	Engineering Manager

Who to survey: Sales team, account managers, strategy team.

11. Team-Level Measurement

11.1 Team Health Indicators

Adapted from the Spotify Squad Health Check Model (Kniberg & Ivarsson, 2012).

Dimension	Green (Awesome)	Amber (Could improve)	Red (Not good)
Delivering value	We deliver stuff users love, on time	We deliver, but not sure users love it	We struggle to deliver the right things
Speed	We get stuff done quickly	Acceptable but could be faster	Constant delays, too much process
Mission	We know exactly why we exist	Roughly clear, could be sharper	No clarity on purpose
Fun	We look forward to coming to work	Work is OK	Dread going to work
Learning	Always learning new things	Learn sometimes	No time to learn
Support	Great organisational support	Could use more support	Blocked or unsupported
Teamwork	We work great together	OK but could be smoother	Silos and blame
Codebase health	Proud of our code	Some areas need work	Afraid to touch things
Product impact	Products make a real difference	Some impactful, others uncertain	Not sure our products matter

11.2 Engineering Satisfaction & Well-being

Indicator	How to Measure	Target
Overtime hours	Hours logged above 50/sprint/person	< 10% of sprints above target
Unplanned work ratio	% of sprint items not planned at start	< 20%
Meeting load	Hours per week in meetings vs focus time	< 30% meetings
PTO utilisation	Are team members taking their leave?	> 90% of entitlement used

11.3 Knowledge Distribution / Bus Factor

Common risk: Many small teams exhibit single-threaded product ownership, where each product has one primary contributor. If that person is unavailable, work on that product stops. Track your team's bus factor using

the template below.

Product	Primary Contributor	% of Product Hours	Bus Factor	Risk Level
[Product A]				
[Product B]				
[Product C]				

Target: Bus factor ≥ 2 for all production products within 6 months.

Mitigation: Cross-training sessions (1 per product per quarter), PR reviews across product boundaries (target: each engineer reviews PRs for ≥ 2 products), documentation currency tracking.

11.4 Innovation Capacity

Category	Definition	Target
New capability	New products or features that didn't exist before	> 50%
Maintenance	Bug fixes, dependency updates, security, infrastructure	20-30%
Support	Unplanned work, incidents, user requests	< 20%

How to track: Categorise sprint items into these three categories at planning time. Track your team's allocation over multiple sprints to identify trends.

11.5 Cognitive Load Trends

Track NASA-TLX across sprints: sustained high load (> 60 for 2+ quarters) = burnout risk. Increasing trend = workload outpacing capacity. Decreasing trend = automation reducing burden (the goal).

12. Invisible Benefits Measurement

These are the “dark matter” of AI ROI: they exist, they have impact, but they are invisible to standard measurement.

12.1 Context Switching Reduction

Each switch costs approximately 23 minutes of recovery time (Mark et al., 2008). Monitoring products and automation workflows eliminate tasks that previously interrupted developer flow. The value is not just the task time saved. It is the switching cost eliminated.

Metric	Method	Target
Switches per day	Self-report diary for 1 week per quarter	< 6/day
Focus blocks per day	Calendar analysis: blocks ≥ 2 hours	≥ 2/day
Daily switching cost (estimated)	Switches x 20 minutes (midpoint estimate)	< 2 hours/day

12.2 Faster Onboarding

Knowledge agents enable new starters to answer their own questions about services and processes. Well-documented products reduce the knowledge transfer burden on existing team members.

Metric	Method	Target
Time to first commit	Git data: days from start to first merged PR	< 5 days
Time to independent work	Manager assessment	< 15 days
Documentation sufficiency	New starter survey (1-5)	> 4.0

12.3 Decision Speed

Context	Before Automation	After Automation	How to Measure
Vendor evaluation	Days-weeks (manual research)	Hours (research tool)	Track from research request to recommendation delivery
Incident response	Hours (manual detection)	Minutes (monitoring system)	MTTD comparison
Proposal review	Days (manual review)	Hours (AI reviewer)	Track from submission to review completion
Customer health assessment	Weekly (manual reporting)	On-demand (reporting tool)	Time from question to answer

12.4 Meeting Load Reduction

Automated reporting tools replace manual report-out meetings. Monitoring alerts replace “anything happening?” check-in meetings. Knowledge agents reduce “Does anyone know...?” messages in chat.

12.5 Knowledge Capture

AI interactions can convert tacit knowledge into searchable, documented procedures. Measure: AI-generated docs (growing), knowledge base size (growing), repeat question rate (decreasing).

> **Knowledge Democratisation Risk**
>
> Users with AI-generated expertise may have false confidence. Require expert validation for high-stakes decisions (Nature, 2024).

12.6 Error/Rework Reduction

Metric	Method	Target
Defects per release	Bug items per deployment	Decreasing trend
Cost per defect	Hours to investigate + fix + verify	Decreasing trend
Rework rate	% of items reopened after closure	< 5%
Customer-reported issues	Support tickets related to automated products	Low and decreasing

13. Implementation Roadmap

Phase 1: Baseline

Month 1

- Deploy first survey: SPACE + eNPS + S-TIAS
- Define “value moments” per product
- Deploy SUS to production product users
- Collect baseline telemetry
- Run team health check
- Baseline NASA-TLX
- Compile baseline report

Phase 2: Instrument

Months 2-3

- Add accept/edit/reject tracking to AI products
- Add CES micro-survey triggers
- Build product health scorecard template
- Add sprint item categorisation
- First measurement cycle
- Review and simplify metrics

Phase 3: Report

Quarterly, Ongoing

- Product health scorecard review
- Team health check (retro)
- SPACE + NASA-TLX survey
- eNPS + S-TIAS per product
- Trend analysis and leadership reporting
- Framework refinement (6-monthly)

Phase 1: Baseline (Month 1)

Week	Activity	Owner
1	Deploy first survey: SPACE (11 items) + eNPS (1 item) + S-TIAS (3 items) to all team members	Engineering Manager
1	Define “value moments” for each product	Product owners
2	Deploy SUS to users of all production products	Engineering Manager
2	Collect baseline telemetry: MAU, usage volume, error rates per product	Product owners
3	Run team health check (Section 11.1) during retro	Engineering Manager
3	Baseline NASA-TLX for the team	Engineering Manager
4	Compile baseline report. Document all scores. Starting point for trend tracking.	Engineering Manager

Phase 2: Instrument (Months 2-3)

Week	Activity	Owner
5-6	Add accept/edit/reject tracking to AI-output products	Product owners
5-6	Add CES micro-survey trigger after key product interactions	Product owners
7-8	Build product health scorecard template (Section 14.1)	Engineering Manager
7-8	Add sprint item categorisation (New/Maintenance/Support) to sprint planning	Engineering Manager
9-10	First measurement cycle: collect all metrics, populate scorecard	All
11-12	Review first cycle. Identify metrics that are difficult to collect or don't add insight. Simplify.	Engineering Manager

Phase 3: Report (Quarterly, Ongoing)

Activity	Cadence	Owner
Product health scorecard review	Quarterly	Engineering Manager + product owners
Team health check	Quarterly (during retro)	Engineering Manager
SPACE + NASA-TLX survey	Quarterly	Engineering Manager
eNPS + S-TIAS per product	Quarterly (staggered)	Engineering Manager
SUS per product	Quarterly (staggered)	Engineering Manager
Trend analysis and reporting to leadership	Quarterly	Engineering Manager
Framework refinement (add/remove/adjust metrics)	Every 6 months	Engineering Manager

Survey Burden Management

A common failure mode for measurement programmes is survey fatigue. The instruments in this framework total approximately 37 items across all surveys. If someone uses 3 products, the theoretical quarterly burden is 48+ questions on top of team-level instruments. This is too much. Apply these principles:

- **Maximum 10 survey items per person per sitting.** Bundle instruments into short sessions.
- **Rotate products through measurement cycles.** Not every product needs every instrument every quarter. Measure 3-4 products per quarter and rotate.
- **Prioritise products in active development or early adoption.** Stable, mature products can shift to 6-monthly measurement.
- **Use micro-surveys at value moments.** A single CES question after a value moment is less intrusive and more informative than a quarterly batch survey.

Survey Calendar (Steady State)

Month	Week 1	Week 2	Week 3	Week 4
M1 (Jan/Apr/Jul/Oct)	SPACE + NASA-TLX (team)	SUS: Group A products	SUS: Group B products	Team health check (retro)
M2 (Feb/May/Aug/Nov)	eNPS + S-TIAS: Group A products	SUS: Group C products	eNPS + S-TIAS: Group B & C products	
M3 (Mar/Jun/Sep/Dec)	Compile quarterly scorecard	Review with team	Present to leadership	Plan adjustments for next quarter

14. Scoring & Reporting

14.1 Product Health Scorecard Template

Product Health Scorecard: [Product Name]

Quarter: [QX YYYY] · Owner: [Product Owner]

DIMENSION	SCORE	STATUS	TREND
Delivery	[XX%]	GREEN	[Trend]
Quality	[XX%]	GREEN	[Trend]
Adoption	[XX%]	AMBER	[Trend]
Satisfaction (eNPS)	[+/-N]	[Status]	[Trend]
Trust (S-TIAS)	[N.N]	[Status]	[Trend]
Cognitive Load	N/A		(team-level)
Strategic Impact	[TBD]	[Status]	[Trend]

KEY METRICS THIS QUARTER

- Units processed: [N]
- Hours saved (validated): [N] hrs/week
- Active users: [N] of [N] eligible ([XX%])
- AI output acceptance rate: [XX%]

TOP ACTION ITEMS

1. [Action item 1]
2. [Action item 2]
3. [Action item 3]

14.2 Team Health Scorecard Template

Team Health Scorecard: [Team Name]

Quarter: [QX YYYY]

DIMENSION	SCORE	STATUS	TREND
Sprint Completion Rate	[XX%]	[Status]	[Trend]
Velocity (items/sprint)	[N] avg	[Status]	[Trend]
SPACE Composite	[N.N]	[Status]	[Trend]
NASA-TLX Composite	[N]	[Status]	[Trend]
Innovation Capacity	[XX%]	[Status]	[Trend]
Bus Factor (avg)	[N.N]	[Status]	[Trend]
Hours Utilisation	[XX%]	[Status]	[Trend]

Products in Production: [N] · In UAT: [N] · In Dev: [N]

Total Hours This Quarter: [N]

KEY ACTIONS

1. [Action item 1]
2. [Action item 2]
3. [Action item 3]

14.3 Quarterly Trend Tracking

Metric	Q-1	Q0 (Baseline)	Q+1	Q+2	Trend
Sprint Completion	--	[XX%]			Baseline
Avg eNPS (all products)	--	[TBD]			Baseline
Avg SUS (all products)	--	[TBD]			Baseline
Avg S-TIAS (AI products)	--	[TBD]			Baseline
SPACE Composite	--	[TBD]			Baseline
NASA-TLX Composite	--	[TBD]			Baseline
Total Hours Saved (validated)	--	[TBD]			Baseline
Innovation Capacity	--	[XX%]			Baseline
Avg Bus Factor	--	[N.N]			Baseline

14.4 Presenting Qualitative Data to Leadership

1. **Lead with the business outcome, not the survey score.** Instead of "Our eNPS is +32," say "8 of 10 users would recommend the product to colleagues, and usage has grown 40% quarter-over-quarter."
2. **Connect qualitative to quantitative.** "Trust scores for the research platform increased from 4.2 to 5.1 this quarter. In the same period, the sales team started using it for 3x more research requests, because they trust the outputs enough to rely on them."
3. **Use the scorecard as a traffic light.** Green/Amber/Red is universally understood. Don't lead with numbers. Lead with colour, then explain the numbers for anyone who wants depth.
4. **Show trends, not snapshots.** Three quarters of improving eNPS means the product is getting better. Three quarters of declining trust means intervention is needed.
5. **Name the "so what?"** Every qualitative metric should connect to a business outcome: trust → adoption → usage → hours saved → cost avoidance or revenue impact. Draw the chain explicitly.

15. Appendices

Appendix A: Full Survey Question Bank

All questions are copy-paste ready for Microsoft Forms, Google Forms, or any survey tool.

A.1 eNPS (1 question per product)

QUESTION

On a scale of 0 to 10, how likely are you to recommend [PRODUCT NAME] to a colleague?

Scale: 0 (Not at all likely) to 10 (Extremely likely)

Follow-up: What is the primary reason for your score? [Free text]

A.2 System Usability Scale (SUS): 10 questions per product

10 QUESTIONS

Scale: 1 (Strongly Disagree) to 5 (Strongly Agree)

1. I think that I would like to use [PRODUCT NAME] frequently.
2. I found [PRODUCT NAME] unnecessarily complex.
3. I thought [PRODUCT NAME] was easy to use.
4. I think that I would need the support of a technical person to use [PRODUCT NAME].
5. I found the various functions in [PRODUCT NAME] were well integrated.
6. I thought there was too much inconsistency in [PRODUCT NAME].
7. I would imagine that most people would learn to use [PRODUCT NAME] very quickly.
8. I found [PRODUCT NAME] very cumbersome to use.
9. I felt very confident using [PRODUCT NAME].
10. I needed to learn a lot of things before I could get going with [PRODUCT NAME].

Scoring: Odd items: response − 1. Even items: 5 − response. SUS = sum × 2.5.

A.3 Customer Effort Score (CES)

QUESTION

"[PRODUCT NAME] made it easy to accomplish my task."

Scale: 1 (Strongly Disagree) to 7 (Strongly Agree)

A.4 S-TIAS: 3 questions per product

3 QUESTIONS

Scale: 1 (Strongly Disagree) to 7 (Strongly Agree)

1. I am confident in [PRODUCT NAME].
2. [PRODUCT NAME] is reliable.
3. I trust [PRODUCT NAME].

A.5 NASA-TLX: 6 questions (team-level)

6 DIMENSIONS

Scale: 0 (Very Low) to 100 (Very High), increments of 5

1. **Mental Demand:** How mentally demanding was your work?
2. **Physical Demand:** How physically demanding was your work?
3. **Temporal Demand:** How hurried or rushed was the pace?
4. **Performance:** How successful were you? *(Low = good, High = poor)*
5. **Effort:** How hard did you have to work?
6. **Frustration:** How insecure, discouraged, irritated, stressed?

A.6 Decision Quality: 5 questions per AI output

5 METRICS

Scale: 1 (Strongly Disagree) to 5 (Strongly Agree)

1. The output was factually correct and free of errors. *(Accuracy)*
2. The output addressed exactly what I needed. *(Relevance)*
3. The output was well-structured and logically organised. *(Coherence)*
4. The output saved me time and effort. *(Helpfulness)*
5. I feel confident acting on this output without further verification. *(Confidence)*

A.7 SPACE Developer Experience: 11 questions

11 ITEMS ACROSS 5 DIMENSIONS

Scale: 1 (Strongly Disagree) to 5 (Strongly Agree)

Satisfaction & Well-being

1. I am satisfied with my work.
2. I feel healthy and energised at work.

Performance

3. I am confident in the quality of my work.
4. I consistently meet the goals I set for myself.

Activity

5. I complete a meaningful amount of work each day.
6. My work contributes directly to team objectives.

Communication & Collaboration

7. My team communicates effectively.
8. I can easily get the information I need from colleagues.

Efficiency & Flow

9. I can maintain focus for extended periods.
10. My tools and processes help rather than hinder my work.
11. I spend most of my time on tasks that require my expertise.

A.8 AI Trust & Confidence Assessment: 5 questions per AI product

5 QUESTIONS

Scale: 1 (Strongly Disagree) to 7 (Strongly Agree)

1. I am confident in [PRODUCT NAME]. *(S-TIAS item 1)*
2. [PRODUCT NAME] is reliable. *(S-TIAS item 2)*
3. I trust [PRODUCT NAME]. *(S-TIAS item 3)*
4. I understand why [PRODUCT NAME] produced the output it did. *(Transparency)*
5. I feel comfortable making decisions based on [PRODUCT NAME]'s output without additional verification. *(Decision confidence)*

A.9 Cognitive Load - Product-Specific: 4 questions

4 QUESTIONS

Scale: 0 (Very Low) to 100 (Very High), increments of 5

1. How mentally demanding is it to use [PRODUCT NAME]?
2. How much effort do you have to put in to get useful results from [PRODUCT NAME]?
3. How frustrated do you get when using [PRODUCT NAME]?
4. How much does [PRODUCT NAME] reduce the mental effort of your overall work? *(Higher = more reduction = better)*

A.10 Overall Product Satisfaction: 3 questions per product

3 QUESTIONS

1. How satisfied are you with [PRODUCT NAME]? Scale: 1 (Very Dissatisfied) to 5 (Very Satisfied)
2. [PRODUCT NAME] is valuable to my work. Scale: 1 (Strongly Disagree) to 5 (Strongly Agree)
3. What one thing would you change about [PRODUCT NAME]? [Free text]

A.11 Responsible AI Assessment: 6 questions per product

QUARTERLY ASSESSMENT

Score each item: Yes / Partial / No

1. **Bias testing:** Have we tested this product's outputs for systematic errors across different input types in the last quarter?
2. **Explainability:** Can a non-technical user understand why this product produced a given output?
3. **Data compliance:** Is the data this product processes stored and retained in accordance with our data governance policies?
4. **Human oversight:** For outputs informing consequential decisions, is human review required before action?
5. **Incident log:** Have we documented any AI failures, hallucinations, or unexpected outputs this quarter?
6. **User consent:** Do users understand they are interacting with an AI system and what data is being processed?

Scoring: Products with 6/6 Yes are operating responsibly. Products with any “No” items require remediation planning. Products with “Partial” items should have improvement targets for the next quarter.

Appendix B: Glossary

Term	Definition
Bus Factor	Minimum team members who would need to leave before a product stalls. Bus factor of 1 = critical single-threaded risk.
CES	Customer Effort Score. Measures how easy it is to accomplish a task. Scale: 1-7.
Cognitive Load	Mental effort required to complete a task. Measured by NASA-TLX. High cognitive load leads to fatigue, errors, and reduced quality.
Context Switching	Moving between unrelated tasks. Each switch costs approximately 23 minutes of recovery time.
Decision Confidence	A user's self-assessed certainty that an AI-generated output is accurate enough to act on.
DORA	DevOps Research and Assessment. Google's software delivery performance programme.
eNPS	Employee Net Promoter Score. % Promoters − % Detractors.
Flow State	Deep, uninterrupted focus. Highest productivity and quality.
Innovation Capacity	The proportion of engineering time allocated to new capabilities versus maintenance and support.
Knowledge Democratisation	Making expertise accessible to non-experts through AI and automation tools.
MAU	Monthly Active Users. Unique users who interact with a product at least once per month.
MITRE AI Maturity Model	A 6-pillar, 5-level framework for assessing organisational AI readiness and maturity.
NASA-TLX	NASA Task Load Index. 6-dimension cognitive load measure. Standard in human factors research.
NIST	National Institute of Standards and Technology. Published the AI Trustworthiness Framework (AI 100-1).
NPS	Net Promoter Score. % Promoters (9-10) minus % Detractors (0-6). Range: −100 to +100.
Override Rate	The percentage of AI-generated outputs that users manually edit or reject.
S-TIAS	Short Trust in Automation Scale. 3-item validated scale for measuring trust in AI/automated systems.

Term	Definition
SPACE	Satisfaction, Performance, Activity, Communication, Efficiency. Microsoft Research's 5-dimension developer productivity framework.
Squishy ROI	UC Berkeley term for soft metrics (adoption, satisfaction, trust) that predict future hard ROI.
SUS	System Usability Scale. Score of 68 = average (50th percentile across 500+ studies).
TEI	Total Economic Impact. Forrester's methodology for calculating the full value of technology investments.
TTV	Time to Value. Duration from first interaction to regular use.
Velocity	Number of parent-level items completed per sprint. Measures delivery throughput, not outcome quality.
Velocity Delta	Completed items minus Planned items per sprint. Target = 0. Positive = over-delivery.

Appendix C: Source References

1. **Anthropic.** "The Anthropic Economic Index." 2025. 57% augmentation / 43% automation split, 1.8% US productivity projection. https://www.anthropic.com/research/the-anthropic-economic-index
2. **Challapally, A., Pease, C., Raskar, R. & Chari, P.** "The GenAI Divide: State of AI in Business 2025." *MIT Project NANDA*, 2025. https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/
3. **Brooke, J.** "SUS: A Retrospective." *Journal of Usability Studies* 8(2), 2013. https://uxpajournal.org/sus-a-retrospective/
4. **Deloitte.** "The AI ROI Paradox." 2024. Deloitte AI Institute (subscription required).
5. **DORA.** "Accelerate State of DevOps Report." Google Cloud, 2024. https://dora.dev/research/
6. **Forsgren, N. et al.** "The SPACE of Developer Productivity." *ACM Queue* 19(1), 2021. https://queue.acm.org/detail.cfm?id=3454124 (subscription required)
7. **Gartner.** "5 Critical Metrics for Measuring AI Value." 2024. Gartner (subscription required).
8. **Gartner.** "AI Maturity Model." 2024. Gartner (subscription required).
9. **GitHub.** "Research: Quantifying GitHub Copilot's Impact on Developer Productivity and Happiness." 2023. https://github.blog/news-insights/research/research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/
10. **Glean.** "Measuring AI Impact: 5 Metrics for AI Decision Quality." 2024. Glean blog (no stable permalink).
11. **Hart, S.G.** "NASA-TLX: 20 Years Later." *Proceedings of the Human Factors and Ergonomics Society* 50(9), 2006. https://humansystems.arc.nasa.gov/groups/tlx/downloads/NASA-TLXtool.pdf
12. **Kaplan, R.S. & Norton, D.P.** *The Balanced Scorecard: Translating Strategy into Action.* Harvard Business Press, 1996.
13. **Kniberg, H. & Ivarsson, A.** "Scaling Agile @ Spotify with Tribes, Squads, Chapters & Guilds." 2012. https://blog.crisp.se/wp-content/uploads/2012/11/SpotifyScaling.pdf
14. **McGrath, M.J., Lack, O., Tisch, J. & Duenser, A.** "Measuring trust in artificial intelligence: validation of an established scale and its short form." *Frontiers in Artificial Intelligence*, 8, 2025. S-TIAS: 3-item short form of Körber's TiA. Cronbach's α = 0.95 to 0.97. DOI: 10.3389/frai.2025.1556737
15. **Mark, G., Gudith, D., & Klocke, U.** "The Cost of Interrupted Work: More Speed and Stress." *Proceedings of CHI*, 2008. https://dl.acm.org/doi/10.1145/1357054.1357072
16. **McKinsey & Company.** "The State of AI in 2024." https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
17. **MITRE.** "AI Maturity Model." 2024. https://aimaturitymodel.mitre.org/

18. **NIST.** “AI Risk Management Framework (AI 100-1).” 2023. https://www.nist.gov/artificial-intelligence
19. **UC Berkeley Haas School of Business.** “Squishy ROI: How Soft Metrics Drive Hard Returns from AI Investments.” 2024. Working paper.
20. **Mahajan, S.** “The democratization dilemma: When everyone is an expert, who do we trust?” *Humanities and Social Sciences Communications*, 2025. DOI: 10.1057/s41599-025-04734-x
21. **Forrester Research.** “The Total Economic Impact of AI-Powered Automation.” 2024. Vendor-commissioned study for Writer. Forrester (subscription required).
22. **Tullis, T. & Stetson, J.** “A Comparison of Questionnaires for Assessing Website Usability.” *Usability Professionals Association Conference*, 2004. Referenced in Section 9 for SUS reliability at small sample sizes.