

Choose the work before the agent

A practical investment framework for deciding where enterprise AI agents belong.

Compare agentic work with simpler alternatives, expose the dependencies behind a promising demonstration, and fund a bounded experiment with a clear decision at its end.

FOR Business sponsors, transformation leaders and product owners

Softlander Consultancy | 2026-09-16 | v1.0

The frameworks and worked scenario are illustrative decision tools, not measured client results or predictions of financial return.

A useful pilot starts with a business choice

An enterprise can build an impressive agent demonstration without resolving whether the work should be delegated to an agent. The missing decision is often upstream: which part of a workflow needs judgment, which part should follow a fixed rule, and which actions deserve human authority? A proposal that begins with the platform tends to inherit the platform's features as its requirements. A proposal that begins with the work can compare automation, assistance, process repair and deliberate inaction on the same terms.

This paper proposes a practical sequence: define a unit of work, establish a credible alternative, separate value from permission, then purchase evidence through a bounded experiment. The result should be a choice among proceed, redesign, defer and stop. None of those choices requires an enterprise-wide maturity score. A team needs enough evidence to decide the next commitment, with the uncertainty and consequence of that commitment made explicit.

The distinction between prescribed workflows and agents that select their own steps is useful here. Anthropic's engineering guidance describes both and recommends starting with simpler designs when they suffice.[1] That supports a question every sponsor can ask: what variation in this work actually requires a model to choose the next action? If the answer is unclear, introduce less autonomy until the need becomes observable.

An agent may be a good candidate when the input varies, relevant evidence is distributed, several tools must be coordinated and a reviewer can judge the resulting work. Those are screening considerations, not eligibility rules. A workflow can satisfy all four and still be unattractive because errors are expensive, access is unavailable, demand is too low, or people have no capacity to absorb the exceptions.

Describe one unit of work precisely

Replace an ambition such as "automate procurement" with a bounded statement: "prepare a comparison of approved suppliers for one purchase request, cite the current contract evidence, and send a draft to the category manager." This description identifies an input, a result, a recipient and an authority boundary. It also exposes what is excluded: selecting the supplier, changing bank information, issuing an order and approving payment.

Capture the current journey from arrival to accepted completion. Include waiting, handoffs, investigation, correction and reopening. A faster first draft is not equivalent to a faster accepted outcome. Ask the operator to walk through an ordinary case and an awkward case using permissible examples. Differences between these cases often reveal the real work more clearly than a process diagram prepared for a steering committee.

Identify the source of truth for every important fact. A supplier's marketing site, an executed contract and an internal eligibility record are different kinds of evidence. Their freshness, access conditions and authority differ. If two sources disagree, decide who resolves the

conflict. Without this rule, a model can produce fluent synthesis while silently making a business-policy decision.

Specify how completion is accepted. "A useful recommendation" is too broad for a pilot. A stronger definition requires the correct supplier set, current contract citations, explicit unresolved differences and no disclosure outside the approved group. Acceptance can combine objective checks with expert judgment, but the expert needs a shared rubric. The same rubric should apply to the current process and the proposed alternative.

NIST's AI Risk Management Framework places context, intended use, measurement and management within an ongoing governance system.[2] For this decision, the practical implication is to keep the workflow, people affected and operating environment together. A model's general capability does not settle the suitability of a particular delegation.

Three decisions before funding a pilot

Separate an attractive opportunity from an authorized, testable workflow.

VALUE	What business outcome changes? Baseline + credible alternative
FEASIBILITY	Can the work be delivered? Inputs + tools + review capacity
PERMISSION	May the action be delegated? Authority + data + accountable owner

Figure 1. Three separate evidence lanes; an unresolved permission cannot be averaged away. Source: original Softlander conceptual framework; no empirical values.

Keep value, feasibility and permission separate

A single opportunity score can conceal a veto. High potential value does not compensate for the absence of lawful access or an accountable decision owner. Equally, a technically easy use case may consume attention without affecting an important business outcome. Review three lanes separately and record the evidence in each.

Value asks what changes for the business or its users. Specify the recurring problem, its frequency, its cost or consequence, the affected population and the current alternative. Distinguish improved experience, released capacity, avoided expenditure and additional revenue. Each needs different evidence. Time returned to a team is useful, but becomes a cash saving only through an actual financial mechanism.

Feasibility asks whether the required work can be performed and operated. Examine representative inputs, integration reliability, policy clarity, review capacity, recovery and measurement. A demonstration usually exercises a narrow path with cooperative systems. An operating design needs a route for incomplete records, conflicting instructions, expired credentials, unavailable tools and an operator who is away.

Permission asks who may delegate which action, using what data, under which conditions. An employee's ability to view a record does not automatically authorize a new automated process to copy, retain or transmit it. Identify the responsible business owner and the security, privacy, procurement or domain specialists whose decisions are needed. Treat unresolved permission as a dependency, not as a low score that can be averaged away.

For each lane, use four evidence states: demonstrated in the target setting, supported by relevant examples, asserted but untested, or unavailable. These labels describe the evidence, not the quality of the team. They prevent a confident slide from looking equivalent to an observed operating result. Record the next question that would move an item to a stronger state.

A compact opportunity brief

Question	Minimum useful answer
What work changes?	One input-to-acceptance journey and its owner
Why change it?	A recurring consequence with a baseline or a plan to establish one
What could replace the agent?	Process repair, ordinary automation, assistance or existing tooling
What may it do?	Permitted actions, excluded actions and escalation boundaries
How will we decide?	Acceptance measures, stop conditions and a named decision date

Do not fill the brief with a score where an answer is missing. "No baseline yet" should lead to measurement work. "No permitted dataset yet" should lead to an access decision or a different test. The brief is effective when it changes the next investment, not when every cell appears complete.

Compare the smallest credible alternatives

Consider four design choices before selecting an agent architecture. First, repair the process: remove duplicate data entry, make a policy explicit or clarify ownership. Second, use deterministic automation for stable rules and known transitions. Third, use AI assistance to gather or draft while a person chooses the next step. Fourth, allow bounded autonomous action where variation and evidence justify it.

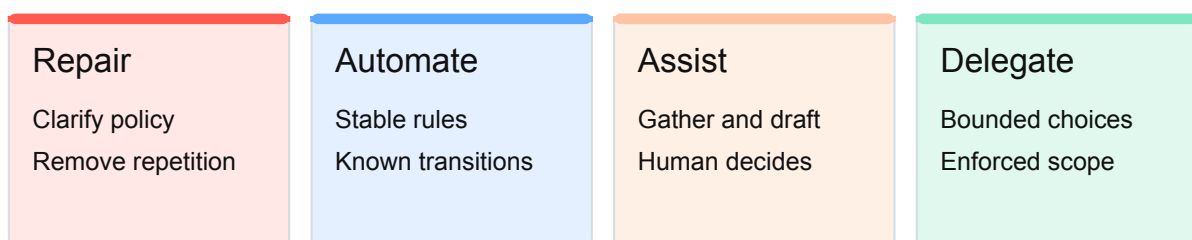
The alternatives can be combined. A request may enter through a conventional form, pass through deterministic eligibility checks, use a model to summarize permitted evidence, and return to a human for a consequential approval. The goal is not maximum autonomy. It is a proportionate allocation of work, with a understandable path from input to accountable action.

Make the comparison at equal service quality. If the agent handles only easy cases while the current process handles all cases, compare the eligible subset and report the excluded work separately. If one option receives expert cleanup and another does not, include that cleanup in both cost and time. Otherwise the apparent gain can come from changing the denominator or moving work out of view.

Published productivity findings illustrate why transfer requires care. A study of a conversational assistant in customer support found that effects differed across workers.[3] That is evidence about a particular deployment and population, not a return estimate for another enterprise. METR's early-2025 study concerned experienced open-source developers; its February 2026 update described selection problems that complicated later estimates.[4][5] These sources justify measuring the local workflow rather than importing a universal productivity assumption.

Choose the least autonomy that serves the work

These options can be combined within one workflow.



More autonomy requires a reason, an authority boundary and relevant evidence.

Figure 2. Compare process repair, automation, assistance and delegation at equal accepted quality. Source: original Softlander conceptual framework; no empirical values.

Design a pilot that can change the decision

Start with the uncertainty that matters most. If the issue is access to evidence, building an elaborate interface answers the wrong question. If the issue is whether managers trust and can review a recommendation, a synthetic output exercise may be more informative than another model comparison. A pilot should make a future commitment less speculative.

Write the decision rule before observing results. Define what would justify expansion, what would trigger redesign and what would stop the work. Avoid choosing a universal pass percentage. The threshold should reflect the consequence of error, the current process and the size of the next exposure. A missing citation and an unauthorized instruction to a payment system should not count as interchangeable defects.

Specify the population of cases and how it is selected. Include ordinary work, difficult variants, missing information and cases that should be declined. Keep deliberately adversarial cases distinct from a representative sample: they answer different questions. The first investigates whether particular failures are possible; the second estimates performance in a defined population, subject to sampling limits.

Run the alternative and the current approach under comparable conditions when feasible. Record accepted completion, review effort, rework, unresolved cases, elapsed time and operating cost. Keep the raw observations needed to understand failures. HM Treasury's evaluation guidance treats the design of evaluation and the explanation of how an intervention produces outcomes as central concerns.[6] Applied here, that means agreeing the causal story before interpreting a dashboard.

Assign authority for stopping the pilot. Operators should not need to negotiate with the project sponsor while an unauthorized action continues. Define containment, human fallback and evidence preservation. These mechanisms are part of the experiment, and testing them may reveal more about readiness than a successful ordinary case.

Illustrative scenario: supplier comparison support

Consider a fictional purchasing team that receives requests for professional services. Its manager wants faster supplier comparisons. The team initially proposes an agent that searches for vendors, ranks proposals and sends selection notices. A workflow review identifies that external search, internal eligibility, negotiation and purchase approval involve different owners. Combining them would make the first experiment difficult to interpret and control.

The team narrows the unit of work to preparing a comparison of three already-approved suppliers for one request. The output must use a current internal contract summary, identify gaps and provide traceable source references. A category manager remains responsible for the recommendation. The system has no tool for sending notices, updating supplier records or creating a purchase order.

The baseline is the current preparation process, including time spent locating contracts and correcting comparisons. The team also tests a simpler alternative: a structured comparison template supported by improved search. If this produces similar accepted quality with less integration and review effort, it may be the better investment. An AI pilot is not entitled to win because it is the newer option.

A permitted case set contains routine requests, incomplete statements of work, conflicting contract dates and a supplier whose eligibility changed. The owner defines serious failures as exposing a restricted commercial term, omitting a required exclusion or presenting an expired term as current. Formatting differences are logged separately. A good average score cannot cancel one of these policy failures.

During the fictional evaluation, imagine that the agent drafts quickly but frequently asks managers to resolve inconsistent source records. That observation would not by itself prove the agent is poor. It might indicate that the organization needs contract-data maintenance before more autonomy. The next funded step would be a targeted evidence and ownership repair, followed by a repeat comparison. No productivity result is assumed in this scenario.

The decision pack contains the eligible workflow, alternative considered, sample definition, accepted outputs, failure examples, human effort and unresolved dependencies. Expansion to sending supplier notices would require a separate authority decision and new tests. Success in preparing a comparison provides no evidence that the system may commit the business externally.

Turn an experiment into an operating commitment

A pilot that passes its local criteria still needs an operating home. Name the service owner, the person responsible for tool credentials, the owner of source quality and the team that receives escalations. These may be different people. Put their responsibilities into the ordinary operating system rather than assuming the pilot team will remain available indefinitely.

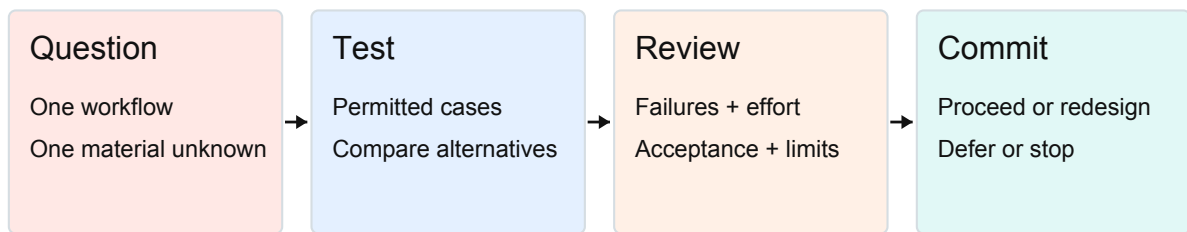
Agree what changes invalidate the evidence. A new model, supplier corpus, tool permission, user group or policy can alter the task even when the interface looks the same. Record the assessed configuration and the events that require reevaluation. NIST's Generative AI Profile identifies risks across the generative AI lifecycle; it is a useful prompt for reviewing changes beyond model output quality.[7]

Plan the exit as carefully as the launch. Retain accessible work records, export essential configuration and identify how users return to the previous process. If the team cannot stop the experiment without losing current work, its commitment is already larger than the sponsor may believe. Portability and recovery deserve attention before a vendor or custom implementation becomes operationally central.

Define what the investment buys next. An assessment may establish a decision and a prioritized backlog. An engineering project may implement an agreed workflow and hand over its tests and operating instructions. Ongoing support may refresh evidence after change. These are different commitments; a positive assessment does not silently authorize implementation or indefinite support.

Fund the next decision, not an open-ended pilot

Each gate changes the size or shape of the next commitment.



Expansion to new actions or users requires a new exposure decision.

Figure 3. Investment gates make stopping and redesigning legitimate outcomes. Source: original Softlander conceptual framework; no empirical values.

Buy a decision or a delivery increment explicitly

The commercial shape of the work should match the uncertainty. If a sponsor cannot yet identify a suitable workflow or judge its feasibility, commission an assessment with a defined decision pack. If the workflow and acceptance criteria are clear, commission an implementation increment. If several teams need repeated prioritization and coordination, an advisory arrangement may be useful. These models solve different problems and should have different acceptance conditions.

For an assessment, ask for the assumptions reviewed, evidence examined, alternatives compared, dependencies and recommended next action. Acceptance should concern the completeness and usability of that work, not a requirement to recommend a build. Otherwise the engagement creates an incentive to turn every investigation into implementation, even when the evidence favors process repair or deferral.

For an implementation increment, define the working artifacts: integration code, configuration, tests, access boundaries, operating instructions and a handover demonstration. Agree what the client supplies, including a business owner, permitted examples, technical access and timely decisions. Distinguish a delay caused by missing access from a defect in the delivered system. The project plan should make both visible without turning every dependency into an open-ended commitment.

For advisory work, specify the decisions and cadence rather than promising unlimited availability. Name the meetings, written outputs and escalation route. Clarify who implements recommendations and who owns delivery outcomes. Advice can help coordinate an operating model, but it does not replace accountable product, engineering, security or business ownership inside the organization.

Procurement should also ask how the work can be continued without the original supplier. Identify which artifacts are handed over, which components have third-party license conditions and which subscriptions the client must maintain. Agree the treatment of confidential data and reusable methods in the engagement terms. A presentation of the final recommendation is not a substitute for usable evidence and an orderly transfer of responsibilities.

These questions should be settled at the level of the next commitment. A small assessment does not need a speculative three-year transformation contract. Equally, a production integration needs more detail than a workshop agenda. The sponsor's objective is to make scope, dependencies and acceptance proportionate to the work being purchased.

Questions for the investment review

Use the following worksheet with a business owner and the people who perform the work. Answers should point to examples or records, not only opinions. A short unresolved answer is preferable to an elaborate invented one.

1. **Work:** What arrives, what leaves, and who accepts completion? Name one concrete case and one legitimate exception.
2. **Baseline:** What quality, human effort, elapsed time and rework are observed today? State where measurement is incomplete.
3. **Alternative:** What is the simplest credible improvement without autonomous action? Explain why it is included or excluded.
4. **Value:** Which business consequence should change, and through what mechanism? Separate capacity, cash, experience and risk.
5. **Authority:** Which actions are permitted, prohibited or approval-dependent? Identify the control that enforces each boundary.
6. **Evidence:** Which cases represent the intended population, and which cases challenge important failure modes? Keep those sets distinguishable.
7. **People:** Who reviews, handles exceptions and owns the resulting work? Confirm their availability rather than assuming it.
8. **Decision:** What results lead to proceed, redesign, defer or stop? Who decides and what commitment follows?

9. **Operations:** Which changes require another evaluation, and how can the workflow be paused or reversed?

10. **Commercial scope:** Which artifacts are being purchased, what constitutes acceptance, and what remains a later decision?

The worksheet is not a certification instrument. It helps a team expose the decisions that a technology demonstration can hide. For highly consequential or regulated uses, specialist legal, security, privacy and domain assessment may change the acceptable design or rule out the proposed delegation entirely.

Make the next commitment smaller and clearer

The best initial opportunity is not necessarily the one with the largest theoretical market or the most impressive demonstration. It is the one where a material business problem, a bounded action space and usable evidence meet. The team should be able to explain why an agent is being considered, what simpler alternative competes with it and what would cause the organization to change course.

Before commissioning a build, assemble a workflow example, an accepted-output definition, an authority map and a short list of unknowns. Use those to scope a decision-focused assessment or a controlled implementation. A clear "not yet" that identifies one repairable dependency can be more valuable than a pilot that ends with enthusiasm but no accountable next step.

For a discussion of a specific workflow, contact [Softlander](#) with the decision you need to make, the current process and the evidence you can use. Do not send confidential records through an initial inquiry. Access, scope and responsibilities should be agreed before detailed analysis begins.

References

- [1] Anthropic. Building effective agents. 19 December 2024. Engineering guidance on workflows, agents and simplicity.
- [2] NIST. Artificial Intelligence Risk Management Framework, version 1.0. January 2023. Voluntary risk-management framework; cited as version 1.0.
- [3] Brynjolfsson, Li and Raymond. Generative AI at Work. NBER Working Paper 31161, April 2023, revised November 2023. Observational deployment study; setting-specific findings.
- [4] METR. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. 10 July 2025. Randomized study in a defined developer population.
- [5] METR. We are Changing our Developer Productivity Experiment Design. 24 February 2026. Follow-up discussing selection and measurement difficulties.
- [6] HM Treasury. Magenta Book: Central Government guidance on evaluation. Current edition accessed 16 September 2026. Evaluation design guidance, not a private-sector requirement.
- [7] NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. July 2024. Cross-sector generative AI risk guidance.