

Artifacts Pack

Every worksheet, scorecard and checklist from the book — one page each, ready to photocopy and fill in by hand.

- | | |
|------------------------------------|---------------------------------|
| 1 · Decision Scorecard | 8 · Measurement Sheet |
| 2 · Vendor Due-Diligence Checklist | 9 · Agent Register |
| 3 · Pre-Mortem Worksheet | 10 · Threat Checklist |
| 4 · Agent Instruction Sheet | 11 · 90-Day Plan |
| 5 · Regulatory Readiness Checklist | 12 · Master Checklist |
| 6 · Case Worksheet | 13 · Final Sign-Off |
| 7 · High-Stakes Gate | 14 · The Rejected-Candidate Log |

How to use this pack. Print the pages you need — or all of them, once. Each page is a single artifact from the book, sized for A4 and designed to be filled in by hand. Tick in ink and date every box: an empty box is a decision you have not yet made, not a task you will get to later.

Decision Scorecard

Score each test 1–5 for the task as precisely scoped. Two and four are the honest middles.

TEST	5	3	1	SCORE
Repetition	Most working days, on a schedule the world controls	Weekly	A handful of times a year	
Mistake-cost	Errors cheap, visible, reversible	Recoverable with real effort and apology	Irreversible harm reaching clients or client money	
Rule-clarity	A competent stranger, trained in an afternoon from your written standard	Written standard exists, with real judgement at the edges	Nobody can state what "correct" looks like	
Regulatory weight	No regulated surface	Regulated records and disclosure, conditions known	The task is itself a regulated decision about people	
TOTAL				

Regulatory weight works differently from the other three: a low score here attaches conditions rather than capping the band — see Override two.

- 1. 16–20: **deploy now.** The machine does the task; a human owns the ends.
- 2. 11–15: **assist-mode only.** The agent drafts, assembles, prepares; a human reviews and sends.
- 3. 10 or below: **don't.** Walk away, or de-scope until it scores.

Override one — the cap. A mistake-cost of 2 or less caps the band at assist-mode, whatever the total. You can satisfy a regulator with conditions; you cannot un-harm a client.

Override two — the regulatory exception. If regulatory weight is the only weak test holding a task below 16, you may deploy — with the conditions named, resourced and logged before go-live.

The asymmetry is not arbitrary: **mistake-cost measures harm you cannot design around; regulatory weight measures conditions you can satisfy.**

Task scored (as scoped):

Band: deploy / assist-mode / don't Date:

Vendor Due-Diligence Checklist

Twelve questions. Ask all of them, in writing, and keep the answers. Questions 1, 2, 3, 8 and 12 are the ones you never skip — the rest can wait for the second meeting.

Vendor:

Meeting date:

1. Which models power it, and are the versions pinned? Why it matters: if the vendor swaps the underlying model mid-contract, your error profile changes overnight. (never skip)
2. What exactly is logged, and can we export it? Logging you can't export is logging you don't really have. (never skip)
3. Who owns our data — and our prompts? Contractual ownership, training opt-outs in writing. (never skip)
4. What is the exit path — in what format, at what cost? Decide the exit before the entrance.
5. Show us the instructions-for-use and the known failure modes.
6. Which sub-processors touch our data?
7. How do you handle disclosure — your AI, our channel?
8. What human oversight does your product assume? Who checks, and when? (never skip)
9. References at our size — may we call them?
10. What does your own eval suite measure, and on what data?
11. Pricing trajectory — what happens after year one?
12. Who on your side is accountable when it's wrong? A name, senior, reachable. (never skip)

Pre-Mortem Worksheet

Do not write how a deployment could fail. Write how YOURS fails — one sentence per box, specific enough to be uncomfortable. A box you cannot fill is not a pass.

1. Automated the ends

The agent reaching past the middle into judgement or ownership.

How *ours* fails:

2. No baseline

Nobody measured the human process, so 'faster' is narration.

How *ours* fails:

3. Bought the demo

Vendor numbers from their harness, never tested on your ugly inputs.

How *ours* fails:

4. No gates

A boundary expressed as a request, not built into the plumbing.

How *ours* fails:

5. No fallback for ambiguity

Nowhere to send 'I can't verify this'.

How *ours* fails:

6. Regulator ignored

Compliance as a launch-day toggle instead of a design input.

How *ours* fails:

7. Agent drift

The agent you deployed is not the agent you have.

How *ours* fails:

Agent Instruction Sheet

One page. If your sheet runs to two pages, you are describing two agents — split it.

Task — One sentence, verb first, scoped by time or count.

Done means — A checkable condition, ideally a number.

Constraints — Scope, hours, sources, tools — written as facts, not hopes.

Must never — The short, absolute list of forbidden actions. Each one also becomes a boundary in the plumbing.

Fallback — What to do when unsure: where to park it, who to ask.

Owner — The named person who reads the trace weekly and answers for the agent — the human the fallback section argued for, now named.

Regulatory Readiness Checklist

One row per item, a who, a when, and the date it was done. The point of the who column is that 'the firm' is not a who; a name is.

#	ITEM	WHO	TARGET	DONE (DATE)
1	Inventory of all AI systems, including shadow AI			
2	Classification: credit scoring / insurance pricing / customer-facing?			
3	Vendor chain papered — instructions-for-use, conformity trajectory			
4	AI-literacy training delivered and recorded			
5	Article 50 disclosure live on customer-facing systems			
6	DPIA / FRIA completed where triggered			
7	Log retention engineered in (6 months; 5 years if transactions)			
8	Workers' representatives notified before deployment (high-risk systems)			
9	Named human owner on file for each agent			

Case Worksheet

One page per candidate task — including the ones you reject. The last row is the one your future self will thank you for.

Task (one line, plain words)	
Score — repetition / mistake-cost / rule-clarity / regulatory weight	_ / _ / _ / _ = __ — band: deploy / assist / don't — cap check: mistake-cost ≤ 2?
Route — borrow / buy / build	the one sentence that justifies it
Controls — instructions	what the agent may do, in describable terms
Controls — boundaries	recipients, systems, amounts — and the approval gate
Controls — fallback	what queues to a human, and who picks it up
Controls — monitoring	the review, its frequency, its owner, log retention
Baseline — hours / cost / quality	measured before deployment, not remembered after
Review date + named owner	a real date and a real person
Rejected because	leave blank only if you haven't been tempted yet

High-Stakes Gate

For any deployment whose blast radius includes client money, credit files or the firm's licence. If any answer is blank, the verdict is already written — and it isn't 'deploy'.

1. What is the worst sentence this agent could produce — and who receives it?

2. Which human reviews and sends, by name — and how fast does their queue move?

3. Where in the output is the AI involvement disclosed?

4. How does a complaint, or a vulnerable client, break out of the agent and reach a person?

5. What is logged, who owns the log, and what will it show when someone official asks?

Measurement Sheet

Eight fields per run; every metric in Chapter 9 derives from them. If it isn't emitted, it isn't measurable.

Deployment:

Task ID	a unique reference, traceable to a specific file or job
Started	when the agent picked the task up
Finished	when it handed it over, done or escalated
Outcome	completed / escalated / failed
Human edits	what a reviewer changed before acceptance — this IS the rework metric
Tokens in / out	the raw fuel cost
Tool calls	what the agent actually did: systems queried, lookups run
Cost	the run's total

Five metrics, weekly:

COMPLETION RATE	REWORK (EDIT DISTANCE)	COST / ACCEPTED TASK	CYCLE TIME	ESCALATION RATE

Agent Register

One page, one row per agent. If you cannot fill in every cell of a row, the agent owns you rather than the other way round.

NAME / TASK	OWNER	TOOLS IT MAY TOUCH	GATES ON ITS ACTIONS	LAST EVAL DATE	NEXT REVIEW	KILL CRITERIA

Monthly: walk the register, read the metrics, log incidents (one page, one hour). Quarterly: re-run the evals, make the kill decisions you have been deferring.

Threat Checklist

One page per agent, six questions. Red-test quarterly and after every update — one hour genuinely trying to make each agent do something forbidden.

- 1. Reads what — every source the agent can retrieve; mark each trusted or untrusted. (Email is untrusted, always.)
- 2. Sends what — every outbound action: emails, file moves, API calls, writes.
- 3. To whom — the allowlist of permitted destinations, and who can edit that list.
- 4. Who could abuse it — outside stranger, client, staff, vendor, and the vendor's next update.
- 5. What stops them — the specific gate, cap, allowlist or human review that blunts each class.
- 6. When last tested — the date of the last red-test, and the date of the last model or vendor update since.

The 90-Day Plan

Four phases, one deployment done properly. Exit criteria are all-three, before you move on.

PHASE	WHAT YOU DO	EXIT CRITERIA
Days 1–15 · Decide	Pick one task; score it; take the baseline; run the pre-mortem; map the regulatory perimeter.	Scored task with a band · measured baseline · mapped perimeter
Days 16–45 · Build & shadow	Buy/build/borrow call; implement the four controls; run in shadow mode — compare daily, then weekly.	Shadow matches human quality on the boring middle · trace clean · zero unexplained gate trips
Days 46–75 · Limited live	One queue or segment; hard caps; approval gates on; log every run; weekly metrics vs baseline.	Completion and rework steady on unfiltered traffic · index card shows the trajectory · caps held
Days 76–90 · Expand or end	Metrics hold: expand one step. They don't: kill or shrink it, publicly, without drama. Either way: runbook, owner, quarterly eval booked.	Decision file complete · runbook written · owner named · eval booked

Standing agenda after day 90: weekly trace review (20 min) · monthly metrics + register (1 hr) · quarterly eval re-run + kill/keep decision.

One-Page Master Checklist

Eleven artifacts, four phases, one page — the whole book as a pre-flight sequence. Photocopy it, one page per deployment, and keep it clipped to that deployment's decision file.

DECIDE — before anything is bought or built

- ☐ Task scored on the Decision Scorecard — with a band
- ☐ Baseline measured — hours, cost, error rate, cycle time
- ☐ Pre-mortem completed — all seven boxes faced
- ☐ Regulatory perimeter mapped — rules, disclosure, named human

BUILD — make it narrow, bounded and visible

- ☐ Route chosen — buy / build / borrow, with vendor diligence if buying
- ☐ Instruction sheet written — task / done / constraints / must-never / fallback / owner
- ☐ Boundaries structural, not advisory — caps, gates, restricted recipients
- ☐ Trace logging on — from the very first run
- ☐ Shadow-mode comparison run — before it touches anything real

PROVE — numbers, unfiltered, before verdicts

- ☐ Measurement Sheet open — every run logged, no cherry-picking
- ☐ Five metrics reported against baseline
- ☐ Steady-state judgement made — 50+ consecutive tasks, unfiltered
- ☐ Index-card unit economics positive — or the deployment shrinks or ends

GOVERN — the deployment you'd still defend in a year

- ☐ Agent Register row created — named owner, boundaries, eval date
- ☐ Quarterly eval booked — against the stored eval set
- ☐ Runbook written — findable by someone other than the builder
- ☐ Threat checklist completed — mitigations verified, not assumed
- ☐ Decommissioning plan noted — logs archived, credentials deleted, row closed

Final Sign-Off

Tick each one. If any box is empty, you haven't finished the plan — go back.

- ☐ Scored task (Decision Scorecard, with a band, not a feeling)
- ☐ Baseline taken (hours, cost, error rate, cycle time)
- ☐ Route decided (buy / build / borrow)
- ☐ Four controls implemented (instruction boundary, approval gate, trace, kill switch)
- ☐ Shadow evidence on file (matching quality on the boring middle)
- ☐ Steady-state metrics against baseline (a month of data)
- ☐ Runbook written (what it does, what normal looks like, three failures + fixes, how to stop it)
- ☐ Named owner in the Agent Register
- ☐ Register row current (scope, caps, owner, last review)
- ☐ Quarterly eval booked (named date, named person)

Deployment:

Owner (name):

Date signed:

The Rejected-Candidate Log

The book's most-used discipline: score the tempting task and reject it in writing, before the pressure arrives. One row per rejection.

CANDIDATE TASK	DATE	WHY REJECTED (SCORES + THE FAILING TEST)	WHO PROPOSED IT

Firms that can say no deploy the yeses with confidence — because confidence isn't the absence of doubt, it's the presence of a recorded reason.