GUIDE G1

TURN8

HOW TO **DEPLOY AI STUDIO AGENTS** INTO LIVE PRODUCTION OPERATIONS

AI Agent Deployment, Integration & Performance Management

WHERE THIS IS USED

- AI Studio agent deployments
- Venture Studio programs with AI-enabled ventures
- Corporate Incubators
- Foundry-as-a-Service engagements
- CVC portfolio companies with AI product components

AUDIENCE

- AI Studio Agent Leads
- Tech Leads
- Venture Studio Program Managers
- EIRs / GMs
- BU Operations Leads receiving the deployment
- Legal and Compliance (for customer-facing agents)

PHASE

Phase Three: Build and Launch → AI Agent Production Go-Live
(Weeks 8–10, following F1 sprint completion)

1

EXECUTIVE SUMMARY

Passing a 50-input test set in a staging environment is not the same as deploying an agent to production. The staging environment is controlled. Production is not. Real users phrase inputs in ways no test set fully anticipates. Concurrent load behaves differently at scale. Data from live interactions exposes edge cases that adversarial test inputs do not. And when something fails in production, it fails in front of customers — not in a controlled test run.

This guide covers the transition from a staging-validated agent to a live production deployment: the go-live readiness checks that must be completed before a single real user touches the agent, the phased rollout strategy that limits blast radius if something goes wrong, the live escalation protocol that is distinct from the testing protocol in Guide B3, the first 48-hour production monitoring window, the data governance and AI compliance checks required for customer-facing deployments, and the rollback decision framework that gives a named individual unconditional authority to restore a stable version within 15 minutes.

The output is an agent that is live, stable, monitored, compliant, and ready for the ongoing performance management cycle in Guide G3.

2

THE CORE PROBLEM

The most common AI deployment failure is not a bad agent — it is a good agent deployed without production infrastructure. The agent performed correctly in staging. It was tested rigorously. But no one mapped the difference between a staging environment with five test inputs and a production environment with five hundred concurrent real users, inconsistent data formats, and a human reviewer who was not briefed on their escalation response obligation.

The failure patterns in production deployment are distinct from staging failures:

- The agent passes staging tests but fails on production data. Real users phrase inputs in natural language that the test set did not capture. The system prompt constraints that worked for 50 curated inputs are too narrow for 500 unpredictable ones.
- Latency is acceptable in staging but exceeds user tolerance in production. API rate limits that never triggered with a single test user fire at concurrent load. The infrastructure was not stress-tested before go-live.
- The escalation protocol was tested in staging but was never activated in production. The human reviewer receives an escalation notification and does not know what to do with it, because no one briefed them before the agent went live.
- Rollback authority is unclear. When the agent fails in production, the team spends 45 minutes identifying who has authority to trigger rollback while the failure accumulates. The rollback protocol existed on paper but was never assigned to a named person with unconditional authority.
- Customer-facing agents go live without AI governance review. In corporate and regulated sector contexts, AI-generated outputs in customer-facing workflows are subject to data localization requirements, sector-specific AI governance standards, and in some cases regulatory notification obligations. Treating governance as a post-launch checklist item creates legal and reputational exposure.
- Logging is configured in staging but not verified in production. The team assumes data from live interactions is being captured. It is not. The logging pipeline was not validated in the production environment before go-live.

The underlying cause in every case is the same:

Production deployment was treated as an extension of the sprint, not as a distinct operational event with its own readiness criteria, authority structure, and monitoring protocol. This guide treats go-live as what it is: the moment a tested, controlled system enters an uncontrolled environment, and the discipline required to do that safely.

3

PREREQUISITES

Must Be Complete Before Starting:

- Guide F1 Sprint Completion Report signed — agent passed the 50-input test set (20 standard, 20 edge, 10 adversarial) at >90% accuracy on standard cases in the staging environment
- Guide E1 confirmed — AI Studio Agent Lead is named, has mandate letter signed, and is 100% allocated
- Guide E3 confirmed — all data access agreements are in place for every data source the agent will access in production
- Guide B3 confirmed — agent specification is documented: system prompt, input/output specification, escalation rule, and success metric for each deployed agent

Production Environment Checks — Must Be Confirmed Before Proceeding:

- Production API keys and credentials are separate from staging credentials — never use staging credentials in production
- Production logging infrastructure is provisioned — logging endpoint, storage destination, and access permissions confirmed
- Agent configuration (system prompt, tools, routing logic) is version-controlled — the exact configuration being deployed is tagged and reproducible
- Rollback version is archived and confirmed deployable — the last stable agent version can be restored in under 15 minutes
- Human reviewer(s) for the live escalation protocol are identified, briefed, and have confirmed their response availability
- For multi-agent or tool-using deployments: cross-agent failure modes mapped, tool authorization scopes confirmed, and shared-state handling documented

4

EXPECTED OUTPUT/ SUCCESS CRITERIA

You have completed this guide when the following are true:

- Agent is live in full production with all users on the production version
- Live escalation protocol is active and has been tested end-to-end in the production environment
- Production logging is confirmed: inputs, outputs, latency, escalation events, and user corrections are all being captured
- First 48-hour production monitoring window completed — no critical failures, rollback threshold not triggered
- For customer-facing agents: AI governance and data localization review completed and sign-off on file before go-live
- Go-Live Report produced and filed — feeds directly into the G3 ongoing monitoring protocol

5

STEP-BY-STEP INSTRUCTIONS

STEP 1 COMPLETE THE GO-LIVE READINESS CHECKLIST

The go-live readiness checklist is a hard-gate document. Every item must be confirmed — not "in progress" or "planned" — before the deployment sequence begins. A partial readiness check is not a readiness check.

Go-Live Readiness Checklist — Four Categories:

CATEGORY	ITEM	CONFIRMED BY	HARD GATE?
Agent Quality	Agent passed 50-input test set at >90% accuracy on standard cases in staging. Zero failures on adversarial inputs. Escalation rate on edge cases < 40%.	AI Studio Agent Lead	Yes — deployment does not proceed without this
Agent Quality	Regression evaluation completed alongside the 50-input test: automated or human-reviewed assessment that confirms the current version performs no worse than the previous version on the existing test set. Configuration version tagged so production behavior is traceable to a specific evaluated version.	AI Studio Agent Lead + Tech Lead	Yes, for any update to an existing agent; strongly recommended for first deployment
Agent Quality	System prompt reviewed and locked — no changes permitted after readiness check until post-deployment review window	AI Studio Agent Lead + Tech Lead	Yes
Infrastructure	Production API keys and credentials configured. Staging credentials removed from production environment.	Tech Lead	Yes — a single staging credential in production is a security failure
Infrastructure	Production logging pipeline verified: 5 synthetic inputs sent, confirmed appearing in the logging system within 60 seconds. Logging captures: inputs, outputs, latency, token usage, tool call traces (if applicable), escalation events, and safety/policy flags.	Tech Lead	Yes — logging not verified = flying blind in production
Infrastructure	Load test completed: concurrent user simulation run at 2x expected peak load. Latency confirmed below 3 seconds at peak. API rate limits confirmed sufficient.	Tech Lead	Yes, for customer-facing; strongly recommended for internal
Infrastructure	Non-functional requirements defined and agreed with the BU before go-live: latency SLA, availability target, error budget, and cost per call / per workflow ceiling. All mapped to monitoring alerts.	Tech Lead + BU Operations Lead	Yes — no go-live without agreed non-functional requirements

CATEGORY	ITEM	CONFIRMED BY	HARD GATE?
Infrastructure	Security posture confirmed: least-privilege access to tools and data sources; secrets managed securely (not hardcoded); prompt injection defenses in place for customer-facing inputs; agent isolated from unrelated systems.	Tech Lead	Yes, for customer-facing; strongly recommended for internal
Infrastructure	Rollback version archived. Confirmed deployable in under 15 minutes. Rollback owner named and briefed.	Tech Lead + AI Studio Agent Lead	Yes — no go-live without a named rollback owner
Escalation	Live escalation protocol activated in production environment — different configuration from staging protocol. Human reviewer(s) briefed on trigger conditions, response time SLA, and what to do.	AI Studio Agent Lead + BU Operations Lead	Yes, for customer-facing; Yes for internal with compliance implications
Escalation	Escalation tested end-to-end in production: trigger condition fired, notification sent, reviewer confirmed receipt	AI Studio Agent Lead	Yes
Governance	For customer-facing agents: AI governance review completed by Legal and Compliance. Data localization requirements confirmed met. Any sector-specific AI standards verified.	Legal / Compliance sign-off	Yes, for customer-facing — hard stop without this
Governance	Data-sharing agreement from E3 confirmed applicable to production data volumes and use cases (not only to pilot-level access)	Tech Lead + Legal	Yes, if data is sourced from corporate parent assets
Governance	Bias and misclassification risk assessment completed for the specific customer segment the agent serves (from the AI Lean Canvas in F1). Actionability and autonomy reviewed: what systems can the agent change, whether changes are reversible, and how agent actions are logged.	AI Studio Agent Lead + Legal	Yes, for customer-facing in regulated sectors; recommended otherwise
Go-Live Plan	Phased rollout plan confirmed: dark launch → shadow mode → canary / partial rollout → full deployment. Stage durations, canary traffic percentages, and gate criteria defined.	AI Studio Agent Lead	Yes

1.1 **Walk every item on the checklist with the AI Studio Agent Lead and Tech Lead** — do not delegate the review — A checklist item that is "confirmed" without being verified is not confirmed. For each item: state the evidence that confirms it, not just the assertion that it is done

1.2 **For any item that is not fully confirmed: stop the deployment clock** — The go-live date moves. The checklist does not. Every open item is assigned a resolution owner and a resolution date. Only when all items are confirmed does the deployment sequence begin

AI PROMPT — Go-Live Readiness Review

```
I am completing the Go-Live Readiness Checklist for an AI Studio agent deployment. Agent name: [name]. Agent function: [describe — customer-facing or internal, what it does]. Production environment: [describe]. For each of the four readiness categories — Agent Quality, Infrastructure, Escalation, Governance — identify any items that are not yet confirmed and classify them as: (1) confirmed with evidence, (2) in progress with a specific resolution date, or (3) not started and blocking go-live. For every blocking item: write the specific action required, name the owner, and set a resolution date. Output as a deployment status report for the AI Studio Agent Lead and the Executive Sponsor.
```

STEP 2 RUN THE AI GOVERNANCE AND COMPLIANCE CHECK

AI governance is not a post-launch audit. For customer-facing agents in corporate and regulated sector contexts, governance clearance is a go-live blocker. An agent that produces AI-generated outputs in customer-facing workflows may be subject to data localization requirements, sector-specific AI regulations, and in some cases regulatory notification or disclosure obligations that vary by market and sector.

Internal operations agents typically carry lower compliance risk and can often go live before customer-facing governance is complete. This is the correct sequencing: deploy the internal agent first, capture real-world data, then deploy the customer-facing agent once governance is cleared.

2.1 Classify each agent by deployment context

- **Customer-facing agents:** Any agent that interacts directly with external users, produces AI-generated outputs in customer communications, or processes personal data of external individuals. These require full governance review before go-live.
- **Internal operations agents:** Any agent that operates within internal workflows, processes only corporate operational data, and does not interact directly with external users. Lower compliance threshold — confirm data governance requirements with Legal but does not require the same pre-launch clearance as customer-facing agents.

2.2 Complete the AI governance review for customer-facing agents

GOVERNANCE DIMENSION	WHAT TO CONFIRM	WHO CONFIRMS	NOTES
Data Localization	Where is production data stored? Where is the API processing the data hosted? Does this comply with any data residency requirement applicable to the customer segment?	Tech Lead + Legal	Confirm data residency requirements with your Legal team for each sector. Cloud provider region selection must be confirmed before go-live, not after. For cross-market deployments: compliance must be confirmed per jurisdiction (e.g., national AI guidelines, sector regulators), not just against corporate policy.
AI-Generated Output Disclosure	Does the user know they are interacting with an AI agent? Are there disclosure requirements in the applicable regulatory framework?	Legal / Compliance	Financial services, healthcare, and HR automation carry higher disclosure requirements. Confirm with your Legal team for each market before deploying customer-facing agents.
Actionability and Autonomy	What systems or data can the agent change or write to? Are those actions reversible if the agent produces an incorrect output? Are all agent actions logged with sufficient detail to explain what happened and why?	AI Studio Agent Lead + Legal	Agents that take actions beyond generating text — updating records, sending messages, triggering workflows — require additional review of reversibility and action logging before go-live.
Security Posture	Are tool and data access permissions scoped to the minimum required? Are secrets and credentials managed securely? Have prompt injection risks been assessed for customer-facing input paths?	Tech Lead + AI Studio Agent Lead	Treat the system prompt and tool permissions as a security boundary. For agents that call external tools or APIs, confirm that a compromised prompt cannot trigger unintended actions in connected systems.
Bias and Misclassificatio n Risk	For the specific customer segment, the agent serves: have the bias risks from the AI Lean Canvas (F1) been assessed? Is there a documented mitigation for each identified risk?	AI Studio Agent Lead + Legal	Particular attention required for agents that make decisions affecting individuals (credit, hiring, access). Document the mitigation for each risk — not just the identification.
Personal Data Processing	Does the agent process, store, or transmit personal data? Is this covered by the data-sharing agreement from E3? Is a Data Processing Agreement required?	Legal / Compliance	Confirm data classification for all inputs the agent receives. If inputs include personal data, a DPA is required before go-live in most regulatory frameworks.
Escalation and Human Override	Is there a documented mechanism for a human to override any AI-generated decision? For regulated outputs: can the affected individual request human review?	AI Studio Agent Lead + Legal	In some regulated sectors, automated decision-making must include a human review option. Confirm whether this applies to the agent function before go-live.

2.3 **Obtain written governance sign-off before deploying any customer-facing agent** — Email confirmation from Legal and Compliance is sufficient. The sign-off is filed in the content library (from E2) alongside the agent specification. The sign-off date is recorded in the Go-Live Report.

STEP 3 CONFIGURE THE PRODUCTION ENVIRONMENT

The production environment is a different system from staging. Not a copy of staging with different credentials — a separately configured, separately monitored, separately governed environment. Every configuration item that was set in staging must be re-confirmed in production, not assumed to have carried over.

3.1 Configure production API credentials and connections

- **Action:** Set production API keys for every external service the agent calls: the LLM API, any data source connections, the workflow tool (n8n, Make, or equivalent), and the notification/escalation channel.
- **Verify:** All staging credentials have been removed from the production configuration. Run a connection test for every integration — not just the primary LLM call. A failing secondary connection (e.g., a CRM lookup the agent uses to enrich outputs) will produce degraded outputs that are harder to diagnose than an outright failure.

3.2 Configure and verify the production logging pipeline

- **What must be logged:** Every input the agent receives (text, format, timestamp, user identifier). Every output the agent produces. Latency per call. Token usage per call (for cost tracking). Tool call traces — which tools were called, with what inputs, and whether they succeeded. Escalation events. Safety or policy flags. User corrections if a feedback mechanism exists.
- **Verification test:** Send 5 synthetic inputs to the production agent. Confirm all 5 appear in the logging system within 60 seconds with complete field capture. If any field is missing or any log entry is absent: stop deployment and resolve before proceeding.

3.3 **Set up the live monitoring dashboard before the first real user interaction** — The monitoring dashboard is configured and verified before go-live — not during the first 48-hour window. Minimum dashboard components:

METRIC	ALERT THRESHOLD	ALERT RECIPIENT	CHECK FREQUENCY
Accuracy rate (human-reviewed sample)	Below 85% on any 50-call sample triggers immediate review	AI Studio Agent Lead	Daily in first week; weekly thereafter
Escalation rate	Above 25% of calls in any 1-hour window triggers immediate review	AI Studio Agent Lead + BU Operations Lead	Continuous — real-time alert
Response latency (p95)	Above 3 seconds for SaaS / above 10 seconds for complex B2B workflow	Tech Lead	Continuous — real-time alert
Error rate (failed calls, not escalations)	Above 1% of calls in any 1-hour window triggers immediate review	Tech Lead	Continuous — real-time alert
Cost per call / rate of spend	Sudden 3x spike vs. forecast baseline — may indicate unbounded loops, runaway tool calls, or unexpectedly long contexts	Tech Lead + AI Studio Agent Lead	Continuous — real-time alert
Traffic volume (calls per hour)	Sudden 5x spike vs. baseline — may indicate abuse, a misconfigured integration loop, or unexpected demand	Tech Lead	Continuous — real-time alert
Safety / policy flags	Any output flagged by content safety filters	AI Studio Agent Lead	Continuous — real-time alert

3.4 **Version-control the agent configuration** — The system prompt, tool list, routing logic, and model version are treated as code: stored in version control, tagged at each deployment, and deployed reproducibly. Rollback is a configuration rollback — restoring a previous version tag — not a manual re-edit of the live system prompt.

3.5 **Assign the rollback owner and confirm the rollback procedure** — The rollback owner has unconditional authority to execute rollback within 15 minutes of identifying a critical failure. No approval chain. No committee. One person, one decision, 15 minutes.

Rollback trigger conditions (any one is sufficient):

- Error rate exceeds 5% of calls in any 30-minute window
- Escalation rate exceeds 50% of calls for more than 15 minutes
- Confirmed production of incorrect outputs on standard cases
- Confirmed data governance breach or personal data exposure
- Cost spike indicating unbounded agent behavior
- Rollback owner assessment: agent behavior is causing user harm or material business impact

Rollback procedure:

- Rollback owner triggers rollback — no approval required
- Previous stable version deployed from archive — target: under 15 minutes
- AI Studio Agent Lead and Tech Lead notified immediately
- Incident log opened with timestamp, trigger condition, and rollback time
- Production traffic directed to the restored stable version or to manual fallback process

AI PROMPT — Production Environment Configuration Verification

```
I am verifying the production environment configuration for an AI Studio agent deployment. Agent: [name and function]. Production platform: [n8n / Make / custom / other]. For each integration the agent uses: [list — LLM API, data sources, CRM, notification channels]. Produce a production configuration verification checklist: (1) for each integration, the specific test that confirms it is working in production (not in staging), (2) the 5-input logging verification test — what inputs to send and what to look for in the logging system, (3) the monitoring dashboard configuration — what to set up, where, and which team member owns each alert, (4) the rollback procedure for this specific agent — the exact steps in sequence with the time target for each. Format as an executable operations runbook, not a summary.
```

STEP 4 EXECUTE THE PHASED ROLLOUT

A phased rollout limits the blast radius of a production failure. If something goes wrong in the dark launch or shadow mode phases, it affects a controlled subset of traffic — not all users simultaneously. Each phase has a defined gate criterion. Gate criteria are assessed against the baseline (the existing process or previous model version) — not just against absolute thresholds. If the gate criterion is not met, the rollout does not advance to the next phase.

PHASE	WHAT HAPPENS	DURATION	GATE CRITERION TO ADVANCE	ROLLBACK IF GATE FAILS
Dark Launch	The production agent runs in the background, processing real inputs but not returning outputs to users. Actual outputs come from the existing process. The agent outputs are logged and reviewed by the AI Studio Agent Lead only.	24–48 hours	Agent accuracy on live inputs > 90% on a reviewed sample of 50+ calls. Escalation rate < 25%. No governance incidents. Cost per call within forecast.	Deactivate dark launch. Review system prompt against failing inputs. Return to staging for prompt revision.
Shadow Mode	The production agent processes real inputs and produces outputs, but outputs are shown alongside the existing process output — not replacing it. The AI Studio Agent Lead reviews for quality. Optionally, use shadow mode to A/B test two candidate prompts or model versions before promoting one.	48–72 hours	User feedback on agent outputs is positive or neutral. Accuracy and cost maintained vs. baseline. Escalation rate < 25%.	Deactivate shadow mode. Agent output removed from user view. Review and revise based on feedback.
Canary Rollout	The agent fully replaces the existing process for a small initial slice of traffic — start at 1–5% and scale gradually (e.g., 1 → 5 → 20 → 50 → 100) based on metrics relative to baseline. Routing is controlled by a feature flag or traffic split configuration, making rollback a quick traffic switch.	48–72 hours per increment	No critical failures at each increment. Escalation rate < 20%. Latency and cost within SLA vs. baseline. User satisfaction neutral or positive.	Rollback owner switches traffic back via feature flag. No users remain on the failed increment.
Full Deployment	Agent serves 100% of users and workflow instances. Full monitoring active. First 48-hour intensive monitoring window begins.	Ongoing	48-hour window with no critical failures, no rollback triggers, all metrics stable vs. baseline.	Full rollback if any rollback trigger condition is met. Incident log opened.

4.1 **Brief the BU Operations Lead before each phase transition** — The BU lead must know what phase the agent is entering, what changes users will experience, who to contact if a problem is observed, and what the rollback plan is. Users should not be surprised by changes to their workflow without advance notice

4.2 **Document the gate criterion result before advancing each phase** — Format: "Phase [name] completed [date]. Gate criterion: [state the criterion]. Result: [met / not met]. Evidence: [specific data — sample size, accuracy rate, escalation rate, any incidents]. Decision: advance to [next phase] / hold for [specific reason] / rollback."

4.3 **Do not compress phase durations to meet a launch deadline** — A launch date does not override a gate criterion. If the dark launch phase produces results that do not meet the advance criterion, the timeline moves — the criterion does not. A phased rollout that skips phases is not a phased rollout

STEP 5 ACTIVATE THE LIVE ESCALATION PROTOCOL

The live escalation protocol is not the same as the escalation tested in Guide B3. The B3 test confirmed that the ESCALATE signal fires and that a notification is sent. The live escalation protocol adds the operational layer: who receives it, what they do with it, in how much time, and what happens to the user experience while the escalation is being resolved.

5.1 Define the escalation path for every agent deployed

ESCALATION TRIGGER	WHO IS NOTIFIED	RESPONSE TIME SLA	WHAT THEY DO	USER EXPERIENCE DURING ESCALATION
Agent outputs ESCALATE: [reason] on any call	Primary human reviewer for this agent (named individual, not a team)	Customer-facing: 15 min. Internal: 1 hour.	Review the specific input. Produce the output manually or confirm a standard fallback. Log the escalation outcome and input pattern for system prompt review.	User receives: "Your request is being reviewed — you will receive a response within [SLA time]." No AI output returned for the escalated call.
Safety or content policy flag triggered (output filtered for harmful, inappropriate, or policy-violating content)	AI Studio Agent Lead	15 minutes	Review the flagged input and output. If the flag is a true positive: update the system prompt constraints. If a false positive: adjust filter sensitivity. Log the incident.	User receives a standard fallback message. The flagged output is never returned. The incident is logged for review.
Tool misuse detected (agent calls a tool outside its expected scope, triggers unexpected system actions, or produces write operations it should not)	AI Studio Agent Lead + Tech Lead	Immediate	Suspend the affected tool permission. Review the triggering input and prompt. Determine whether this is a prompt injection attempt or a logic gap in the system prompt. Do not restore tool access until root cause is confirmed.	Agent deactivated for tool-dependent calls until the issue is resolved. Manual process activated.
Error rate alert triggered (> 1% of calls fail)	Tech Lead	15 minutes	Review error logs. Identify whether infrastructure (API failure, rate limit) or agent logic. Resolve or trigger rollback accordingly.	If widespread failure: activate manual fallback. Do not leave users receiving error messages without an alternative path.
Latency alert triggered (p95 > 3 seconds)	Tech Lead	30 minutes	Identify latency source: LLM API, tool calls, or output processing. Check rate limits and API tier.	User experience degrades gracefully — display a loading indicator. Do not time out silently.
Governance incident (suspected data breach, personal data exposure, or AI output causing user harm)	AI Studio Agent Lead + Legal + Tech Lead	Immediate	Rollback owner triggers rollback immediately. Legal notified. Incident log opened with full capture of the triggering input and output.	Agent deactivated immediately. Users notified the service is temporarily unavailable.

5.2 **Test the live escalation protocol in the production environment before dark launch —** Send a synthetic input designed to trigger ESCALATE: in the production agent. Confirm: the notification fires, reaches the correct reviewer, the reviewer confirms receipt, and the response time is within SLA. If any step fails: do not proceed to dark launch.

5.3 **Brief every escalation reviewer before go-live with a written protocol sheet —** One page per reviewer: the specific trigger conditions that will fire their notification, the expected response time, the exact steps to take, the fallback if they cannot respond, and who to call for help. For high-criticality customer-facing agents, define on-call rotations with named backup reviewers and confirmed coverage across business hours.

STEP 6 RUN THE FIRST 48-HOUR PRODUCTION MONITORING WINDOW

The first 48 hours of full production deployment are the highest-risk window. Production inputs are unpredictable in ways that no test set fully captures. Concurrent load may behave differently than anticipated. Edge cases that appeared at low frequency in shadow mode may appear at high frequency in full deployment. The first 48 hours require heightened monitoring attention — not a handover to automated alerts alone.

6.1 **Assign the AI Studio Agent Lead and Tech Lead to active monitoring for the first 48 hours** — This means checking the monitoring dashboard at minimum every 2 hours, not relying solely on automated alerts. Automated alerts catch threshold breaches. Human monitoring catches pattern shifts that have not yet reached an alert threshold.

6.2 **Review a human-evaluated sample of 50 production outputs at the 24-hour mark** — Do not rely solely on automated accuracy metrics in the first 48 hours. The accuracy calculation is only as good as the automated evaluation logic, which may not catch subtle output degradation. A human reviewer evaluates 50 randomly selected production outputs against the success criteria from the B3 agent specification.

6.3 Apply the 48-hour monitoring decision framework

48-Hour Signal	Assessment	Action
All metrics within threshold. Human sample review confirms outputs are accurate and relevant. No escalation spikes.	Healthy production deployment.	Proceed to standard weekly monitoring cadence (Guide G3). Produce the Go-Live Report. Brief the Venture Board in the next monthly milestone review.
Metrics within threshold but human sample review identifies subtle output quality issues not captured by automated metrics.	Production inputs are exposing edge cases the test set did not cover. The agent is functioning but not optimally.	Do not rollback. Capture the specific input patterns causing quality issues. Schedule a system prompt revision for within 72 hours. Monitor closely. Document as a known limitation in the Go-Live Report.
Escalation rate trending upward. Latency degrading slightly. No single alert threshold breached yet.	Early warning signals. Not yet a rollback trigger but requires immediate investigation.	AI Studio Agent Lead and Tech Lead hold a 30-minute review. Identify root cause. If the cause is identified and a fix is available within 4 hours: apply it. If not: rollback owner makes the call on whether to rollback or maintain with increased monitoring.
Any rollback trigger condition met: error rate > 5%, escalation rate > 50% for 15+ minutes, confirmed incorrect outputs on standard cases, governance incident.	Critical failure. Production deployment is not stable.	Rollback owner executes rollback. Target: under 15 minutes. Incident log opened. Rollback confirmed to Executive Sponsor within 1 hour. Post-rollback review within 24 hours.

6.4 For any rollback or significant incident: hold a blameless post-mortem within 72 hours — The post-mortem produces three documented outputs: the specific root cause (not "the agent failed" — the specific input, configuration, or infrastructure condition that caused the failure), the specific change to the test set, system prompt, or infrastructure that addresses it, and the revised deployment timeline. Every failing input from a production incident becomes a test case in the evaluation pipeline before re-deployment.

6.5 Produce the Go-Live Report within 24 hours of completing the 48-hour window

Go-Live Report — Required Content:

- Agent name, deployment date, deployment context (customer-facing or internal), production platform
- Readiness checklist completion status — all items confirmed before go-live
- Governance clearance status — sign-off date and approver name for customer-facing agents
- Phased rollout summary — each phase, gate criterion result, and duration
- 48-hour monitoring window summary — key metrics, any incidents, any known limitations identified
- Live escalation protocol status — any escalations triggered, response time achieved, issues identified
- Production logging confirmation — logging verified, data capture confirmed for all required fields
- Current agent status: Stable (no issues) / Monitored (minor issues being addressed) / Rolled Back (with reason)
- Next review date — first weekly performance review per Guide G3 scheduled

TROUBLESHOOTING

SYMPTOM	LIKELY CAUSE	FIX
Agent passes all 50 test cases in staging but produces wrong outputs in production	Production data is structurally different from the test dataset — real users phrase inputs in ways the test set did not anticipate	Immediately activate the rollback protocol. Do not attempt to fix system prompts in production. Restore the last stable version. Capture a sample of the failing production inputs. Expand the test set using these real-world inputs. Re-test in staging before any re-deployment
Cost per call or total spend spikes significantly above forecast	Unbounded loops, unexpectedly long contexts, or runaway tool calls — the agent is consuming far more tokens or making far more API calls than expected	Activate the cost alert. Identify whether the cause is a specific input pattern triggering a loop or a configuration error. Add explicit context length limits and tool-call count limits to the system prompt. Investigate before restoring full traffic.
Unusual input patterns detected — high volume from a small number of sources, or inputs designed to override system prompt instructions	Potential prompt injection attempt, abuse, or a misconfigured integration sending repeated requests	Do not attempt to patch the prompt in production. Activate tool suspension for any tool that could be misused. Review the inputs in the log for signs of injection: instructions to "ignore previous instructions," requests for system prompt content, or attempts to redirect the agent to unintended actions. Escalate to Tech Lead and AI Studio Agent Lead immediately.
Agent behavior changes unexpectedly after an underlying model update from the API provider	The LLM API provider updated the model version silently — behavior drifted without any change to the system prompt or configuration	This is a model supply-chain risk. Run the regression evaluation against the new model version immediately. If the results regress, pin the previous model version in the API configuration. Add model version pinning to the deployment checklist for all future go-lives.
Escalation rate spikes above 40% in the first 48 hours of production	Edge case definitions are too narrow — the agent is treating valid inputs as edge cases and escalating instead of responding	Widen the acceptable input definitions in the system prompt. Run shadow mode for an additional 24 hours before re-deploying to full traffic. Review the escalation log: are these inputs genuinely ambiguous, or is the agent being overly conservative?
Latency is acceptable in staging but exceeds 3 seconds per call in production	Concurrent user load in production exceeds what staging simulated — the API rate limits or infrastructure did not scale	Review API rate limits and request throttling configuration. Implement request queuing. If using an off-the-shelf API, check the tier and upgrade if necessary. Add latency monitoring with a 3-second alert threshold to the live dashboard
The rollback decision is delayed because no one has authority to trigger it	Rollback authority was not assigned before go-live. The team is waiting for approvals during a live production failure	The rollback owner must be named and confirmed before go-live (Step 3). They have unconditional authority to execute rollback within 15 minutes of identifying a critical failure. Approval chains have no place in a rollback protocol
Data from production interactions cannot be accessed for performance analysis	Logging was configured in staging but not verified in the production environment	Verify the logging pipeline end-to-end in the production environment before go-live, not after. Run a synthetic test in production: send 5 known inputs, confirm they appear correctly in the logging system within 60 seconds
AI governance or data localization review was not completed before customer-facing deployment	The team treated governance as a post-launch checklist item rather than a go-live blocker	This is a hard stop. Customer-facing agents with AI-generated outputs require Legal and Compliance sign-off before deployment in regulated contexts. Deactivate the customer-facing agent immediately. Deploy the internal operations agent only until governance approval is complete. Resume the customer-facing deployment process from Step 2.
Shadow mode produces clean results but full deployment produces unexpected failures	Shadow mode ran on a representative sample, but production traffic has higher volume and greater input diversity than the sample captured	Extend shadow mode to capture at least 500 real production inputs before switching to full deployment. Review the shadow logs for any input patterns that were not in the test set. Do not treat a clean 100-input shadow mode as equivalent to a clean 500-input shadow mode
The live escalation protocol fires but the human reviewer does not respond within SLA	The escalation routing was tested in staging, but the human reviewer was not briefed on their response obligation in production	Brief all escalation reviewers before go-live with: the escalation trigger conditions, the expected response time (default: 15 minutes for customer-facing, 1 hour for internal), and what to do if they cannot respond. Name a backup reviewer for every primary escalation role

7

VALIDATION STEPS

Confirm each of the following before declaring the production deployment complete and stable:

- [ ] Go-Live Readiness Checklist — all four categories confirmed with evidence, not assertion
- [ ] For customer-facing agents: AI governance and data localization review completed with written sign-off on file
- [ ] Production logging verified with 5-input synthetic test — all fields captured within 60 seconds
- [ ] Live escalation protocol tested end-to-end in production — trigger confirmed, notification confirmed, reviewer response confirmed
- [ ] Rollback version confirmed deployable in under 15 minutes — rollback owner named, briefed, and confirmed
- [ ] Phased rollout completed with documented gate criterion results at each phase transition
- [ ] 48-hour production monitoring window completed with no rollback triggers
- [ ] Human sample review (50 outputs) at the 24-hour mark completed with no critical quality issues
- [ ] Go-Live Report produced and filed in the content library (E2)
- [ ] Guide G3 first weekly monitoring review scheduled

8 NEXT STEPS

Upon completing this guide with a stable production deployment:

- **GUIDE G2** — How to Enable Users and Hand Over AI Agent Operations to the Business (run in parallel with G3 from Day 1 of stable deployment)
- **GUIDE G3** — How to Monitor AI Agent Performance, Detect Drift, and Report to the Venture Board (weekly monitoring cadence begins immediately after the Go-Live Report is filed)

If the deployment results in a rollback: the rollback is not a failure — it is the production safety mechanism working correctly. After a rollback, the post-rollback review produces three outputs: the specific root cause of the failure, the specific system prompt or infrastructure change required to address it, and the revised timeline for re-deployment. The phased rollout restarts from dark launch, not from the phase at which the rollback occurred.

MASTER CHECKLIST

A. GO-LIVE READINESS

☐ Agent Quality: 50-input test passed at >90% accuracy on standard cases, zero failures on adversarial inputs, escalation rate < 40% on edge cases

☐ Agent Quality: regression evaluation completed — current version performs no worse than previous version on the test set; configuration version tagged

☐ Agent Quality: system prompt locked — no changes permitted after readiness check sign-off

☐ Infrastructure: production API keys configured, staging credentials removed, all connections tested individually

☐ Infrastructure: logging pipeline verified with 5-input synthetic test — all fields captured: inputs, outputs, latency, token usage, tool call traces, escalation events, safety flags

☐ Infrastructure: load test at 2x expected peak load completed — latency and cost within SLA at peak

☐ Infrastructure: non-functional requirements defined and agreed with BU — latency SLA, availability target, cost ceiling per call / per workflow

☐ Infrastructure: security posture confirmed — least-privilege tool access, secrets managed securely, prompt injection defenses in place

☐ Infrastructure: agent configuration (system prompt, tools, model version) version-controlled and deployable reproducibly

☐ Infrastructure: rollback version archived and confirmed deployable in under 15 minutes

☐ Infrastructure: rollback owner named, briefed, and confirmed available

☐ Escalation: live escalation protocol configured in production (not staging configuration)

☐ Escalation: safety/content policy filter active and tested

☐ Escalation: tool misuse escalation path defined and tested

☐ Escalation: human reviewer(s) briefed with written protocol sheet before go-live; on-call backup named for high-criticality agents

☐ Escalation: end-to-end escalation test in production completed — trigger, notification, and reviewer response all confirmed

☐ Governance: customer-facing agents — AI governance review completed covering data localization, disclosure, actionability/autonomy, security posture, bias risk, personal data, and human override

☐ Governance: data-sharing agreements from E3 confirmed applicable to production data volumes

☐ Governance: cross-market compliance confirmed per jurisdiction where applicable — not just corporate policy

B. PRODUCTION ENVIRONMENT CONFIGURATION

☐ Production environment is a separate, independently configured system — not a copy of staging

☐ All integrations tested individually in production: LLM API, data source connections, tool connections, workflow tool, notification channels

☐ Monitoring dashboard configured and verified before first real user interaction

☐ Seven monitoring metrics active with correct alert thresholds: accuracy, escalation rate, latency (p95), error rate, cost per call / spend rate, traffic volume, safety flags

☐ Alert recipients confirmed for each metric — named individuals, not team mailboxes

☐ Cost alerts configured with a rate-of-spend ceiling to catch unbounded loops or runaway tool calls

☐ Rollback trigger conditions documented including cost spike condition; rollback owner confirmed with unconditional authority

☐ Agent configuration version-controlled — rollback is a version restore, not a manual re-edit

C. PHASED ROLLOUT

☐ Dark launch phase completed — 24–48 hours, gate criterion met (>90% accuracy on reviewed sample, escalation rate < 25%, cost within forecast)

☐ Shadow mode phase completed — 48–72 hours, gate criterion met vs. baseline (accuracy maintained, escalation rate < 25%, user feedback neutral or positive)

C ☐ anary rollout phase completed — traffic incremented gradually (1 → 5 → 20 → 50 → 100%), gate criterion met at each increment vs. baseline

☐ Gate criterion result documented at each phase transition before advancing

Routing controlled via feature flag or traffic split — roll back is a traffic switch, not a re-deployment

☐ BU Operations Lead briefed before each phase transition

☐ No phase duration compressed to meet a launch deadline

D. FIRST 48-HOUR PRODUCTION MONITORING WINDOW

☐ AI Studio Agent Lead and Tech Lead assigned to active monitoring — dashboard checked every 2 hours minimum

☐ Human-evaluated sample of 50 production outputs reviewed at 24-hour mark

☐ 48-hour monitoring decision framework applied — status assessed as Stable, Monitored, or Rolled Back

☐ All monitoring metrics within threshold at 48-hour mark, including cost per call

☐ Any incidents, escalations, or known limitations documented

☐ For any rollback or significant incident: blameless post-mortem held within 72 hours with root cause, test set update, and revised deployment timeline

☐ Go-Live Report produced within 24 hours of completing the 48-hour window

☐ Go-Live Report filed in the content library (E2)

☐ Guide G3 first weekly monitoring review scheduled