

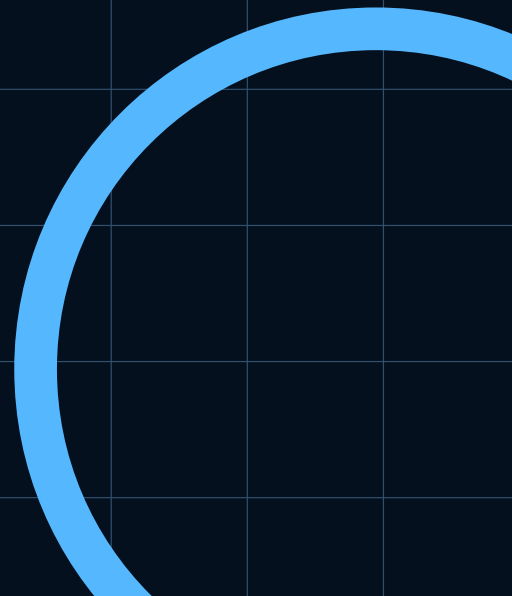
BIGCIRCLE / AI WITH WRITE ACCESS

The Agent Action Boundary Canvas

Define the smallest authority that can complete one real action reliably.

One workflow. One system boundary. One state-changing action.

A working document for product, engineering, security, compliance, and the workflow owner.



How to use the Canvas

Bring these owners into one working session:

1. Business or operational owner
2. Product or workflow owner
3. Engineering or AI owner
4. Security, risk, or compliance owner

Complete each field with evidence.

If the team cannot give a precise answer, mark the field UNRESOLVED.

Do not replace a missing boundary with words such as “appropriate,” “normal,” “as needed,” or “authorized users.”

At the end, select one authority level for this action class:

- **Level 1. Read:** Retrieve permitted information. Do not change state.
- **Level 2. Recommend:** Propose one action with evidence. A person decides and acts.
- **Level 3. Prepare:** Create the exact pending change. A person reviews and submits it.
- **Level 4. Act after approval:** Execute one exact action after fresh authorization.
- **Level 5. Act inside a bounded policy:** Execute without case-by-case approval inside proven limits.

The decision applies only to the action class defined here. It does not grant authority to the full workflow.

Part 1: Workflow and system

Workflow

Workflow name: _____

Business outcome: _____

Trigger: _____

System of record: _____

Other systems touched: _____

Business owner: _____

Operational owner: _____

Current process

What does a person do today from the trigger to the final recorded state?

Which decisions, exceptions, and approvals remain human-owned?

What final state proves that the workflow is complete?

Part 2: Exact state-changing action

Do not write “manage the account,” “handle the ticket,” “process the case,” or “update the CRM.”

Write the exact change.

Action verb: _____

Target system: _____

Target record or resource type: _____

Eligible records or resources: _____

Allowed fields or resource properties: _____

Allowed parameter range: _____

Allowed environment or tenant: _____

Maximum volume or frequency: _____

Actions always denied: _____

Value and consequence

Value if completed correctly: _____

Worst credible result if wrong: _____

Can a person detect the wrong result quickly? Yes / Partly / No

Is the action reversible? Yes / Partly / No

Rollback or compensating action: _____

Part 3: Principal, identity, and delegation

Authentication says who the agent is.

Authorization says what it may do.

Delegation says why it may perform this action now, for this principal, on this task.

Delegating principal

Human or business principal: _____

Principal role: _____

Purpose of the delegation: _____

Request or case identifier: _____

Agent identity

Agent identity: _____

Agent version: _____

Policy version: _____

Tenant and environment: _____

Delegation lifetime

Delegation starts: _____

Delegation expires: _____

Condition that ends it early: _____

Revocation owner: _____

Part 4: Data, tool, and permission boundary

Read boundary

Data the agent can read: _____

Data the agent cannot read: _____

Required source or provenance: _____

Tool boundary

Tools the agent can discover: _____

Tools eligible for this workflow: _____

Tools the agent can invoke for this task: _____

Tools always denied: _____

Action boundary

Actions allowed without approval: _____

Actions that require approval: _____

Actions always denied: _____

Discovery does not grant authority.

The policy must check the agent, principal, task, target, parameters, environment, and current state before it permits the tool call.

Part 5: Credential boundary

Credential source: _____

Credential type: _____

Credential scope: _____

Credential lifetime: _____

Credential revocation path: _____

Can the agent retain or export the permanent secret? Yes / No

If yes, stop. Record why permanent possession is required and who accepted that risk.

The preferred pattern is task or session access. The agent can use the approved capability without owning a permanent secret.

Part 6: Evidence and approval before action

Evidence gate

What must be true before the action can be submitted?

- **Required input:** _____
- **Required provenance:** _____
- **State precondition:** _____
- **Rule or confidence threshold:** _____
- **Policy decision:** _____
- **Missing evidence behavior:** Stop / Ask / Escalate

Approval rule

Named approver or role: _____

Approval applies to: One action / Action class / Time window

Approval expires: _____

Can the target or parameters change after approval? Yes / No

If yes, require a new decision.

Show the approver:

- the exact action;
- target and parameters;
- current state;
- evidence and provenance;
- expected result;
- worst credible failure;
- reversal or recovery path;
- credential and approval expiry.

Stop conditions

The agent must stop when:

- required evidence is missing;
- current state differs from the expected precondition;
- the action exceeds a field, amount, volume, tenant, or environment limit;
- authorization, approval, or credential has expired;
- the tool response is uncertain;
- the first action may have occurred but cannot be confirmed;
- a retry, time, or cost limit is reached;
- the action enters a forbidden state;
- the system cannot prove a safe next action.

Additional stop condition: _____

Part 7: Execution and final-state verification

Expected-state contract

State before action: _____

Expected state after action: _____

Fields or resources expected to change: _____

Invariants that must remain true: _____

Expected downstream effects: _____

Independent verification

The agent must not be the only system that decides whether its own action worked.

Independent verification source: _____

Read-after-write query or check: _____

Correlation identifier: _____

Evidence retained: _____

Select one result state:

- ☐ Verified complete
- ☐ Incomplete and safe to resume
- ☐ Partly complete and needs recovery
- ☐ Uncertain and needs investigation

`Tool` returned `success` is an execution event.

The `intended state exists` and the `invariants hold` is task completion.

Part 8: Idempotency, retry, and partial failure

Idempotency

Business-operation idempotency key: _____

Where execution state is stored: _____

States used: planned → authorized → submitted → confirmed → verified

Retry

Maximum attempts: _____

Maximum time or cost: _____

Reconciliation step before retry: _____

Evidence that proves the first action did not occur: _____

Do not retry an unknown result until the system of record shows that the action did not occur.

Partial failure

Last verified checkpoint: _____

Safe resume rule: _____

Recovery owner: _____

Compensating action: _____

Customer or operator notification rule: _____

A retry repeats the same safe operation.

Recovery handles partial state. It can resume from a checkpoint, complete one missing step, compensate, or stop with evidence.

Part 9: Audit and containment

Audit record

For each state-changing action, retain:

- human or business principal;
- agent identity and version;
- delegation purpose, scope, and expiry;
- user instruction or workflow trigger;
- policy version and authorization decision;
- evidence and provenance used;
- tools considered and selected;
- approval event;
- credential reference and expiry, never the secret;
- exact target and parameters;
- tool request, correlation ID, and raw result;
- final-state verification;
- retry, recovery, and compensating actions;
- containment events.

Audit owner: _____

Retention period: _____

Investigation path: _____

Containment

Kill-switch owner: _____

Immediate containment action: _____

Credential revocation method: _____

Affected-state discovery method: _____

Recovery communication owner: _____

Part 10: Authority decision

Selected level for this action class

- ☐ Level 1: Read
- ☐ Level 2: Recommend
- ☐ Level 3: Prepare
- ☐ Level 4: Act after approval
- ☐ Level 5: Act inside a bounded policy
- ☐ Do not automate this action

Why this is the smallest useful authority

Evidence required for the next promotion

Task-completion evidence: _____

High-impact failure limit: _____

Human-review limit: _____

Required retry and recovery test: _____

Required security or compliance evidence: _____

Decision owner: _____

Decision date: _____

Review or expiry date: _____

Change one material authority property at a time. Do not widen the tool set, data scope, financial limit, and approval policy in one release.

Decision rules

Keep read-only when

- the target action is not precise;
- the principal or agent identity is not attributable;
- the data or tool scope is broad and unresolved;
- the final state cannot be verified independently;
- no owner can contain or investigate the action.

Use recommend or prepare when

- the output is useful but the final decision must remain human-owned;
- the agent can assemble evidence and exact parameters;
- a person must still judge an exception, legal duty, clinical effect, or customer consequence.

Require approval when

- the action affects customers, money, access, compliance, or production;
- the action is difficult to reverse;
- the action class is stable but the failure cost remains high;
- the business owner must retain accountability for each decision.

Permit bounded execution when

- the action class is precise and valuable;
- records, fields, parameters, tools, tenant, environment, and expiry are fixed;
- the system verifies the final state independently;
- retries are idempotent;
- partial failures have a tested recovery path;
- the audit trail preserves the authority chain;
- a named owner can revoke access and contain the agent.

Do not automate when

- the action is not valuable enough to justify the control layer;
- wrong actions are high-impact, hard to detect, and hard to reverse;
- required judgment cannot be represented with stable evidence or rules;
- nobody owns the outcome, recovery, or incident response.

Final question

What is the smallest amount of authority that can complete this task reliably?

Start there.

Expand only when the current boundary produces the evidence required for the next permission.