OPENAI ARCHITECTURE PRIMER

# Codex GitHub Action

## A six-control CI trust boundary

A visual guide to trigger trust, prompt provenance, runner privilege, Codex permissions, model credentials, and GitHub write authority.

FIRST-PARTY DOCUMENTATION / INDEPENDENT LUMINITY SYNTHESIS



# The Codex Action Is a Six-Control CI Trust Boundary

One workflow crosses actor, content, runner, Codex, credential, and GitHub authorization boundaries.

SIX CONTROLS

**01**
**Trigger**
Who may start it

**02**
**Prompt**
What enters context

**03**
**Runner**
Host privilege

**04**
**Codex**
Filesystem + network

**Model credential** 05

Protect the API key and minimize the runtime that can access it.

**GitHub authority** 06

Keep repository write permission in a separately authorized job whenever the workflow permits it.

IN THIS PRIMER

## Key Points

**01 Build the workflow boundary**

Fix the trigger, checkout, prompt sources, permission profile, versions, and output contract.

**02 Govern the blast radius**

Keep actor trust, content trust, host privilege, Codex permissions, and tokens distinct.

**03 Separate consequence**

Generate a reviewable result, then apply only through a narrow downstream identity.

The six-control trust boundary is Luminity synthesis. OpenAI documents the individual Action inputs, safeguards, and workflow pattern.



© 2026 Luminity Digital, Inc. | Independent synthesis of official OpenAI documentation.

Source: https://learn.chatgpt.com/docs/github-action · verified Nov 5, 2026

OpenAI Select Partner

# Build the Workflow Boundary

The workflow fixes what starts the run, what is checked out, what enters the prompt, and what executes.

01

## Which inputs are trusted?

| | |
|---|---|
| ✓ Workflow configuration | REVIEWED |
| ✓ Explicit checkout revision | RECORDED |
| ✓ Prompt template | CONTROLLED |
| ! PR text and repository files | UNTRUSTED |
| ! Fetched external content | UNTRUSTED |

Trusted config · Untrusted content · One prompt

02

## How the run is assembled

Trigger · actor + event | Versions · Action + CLI | Output · file or message

1 **Authorize** — actor write access

2 **Checkout** — explicit revision

3 **Configure** — prompt + permissions

4 **Execute** — codex exec

**ACTOR**
**Default check requires repository write access**

**CONTENT**
**Trusted actor does not make content trusted**

**RESULT**
**Final message, file, or structured output**

03

## Four workflow decisions

### Checkout target — SOURCE

Use an explicit reviewed ref and disable persisted checkout credentials.

### Prompt provenance — INPUT

Identify every event field, repository file, and external source that reaches the model.

### Permission profile — CODEX

Prefer the narrowest supported profile; keep runner safety as a separate decision.

### Resolved versions — DRIFT

Pin the Action commit and CLI release, then record what the runner actually executed.

**BOUNDARY**

- Trusted trigger ≠ trusted content
- Workflow config is executable policy
- Permission profiles ≠ runner safety
- Output remains untrusted



© 2026 Luminity Digital, Inc. | Independent synthesis of official OpenAI documentation.

Source: https://learn.chatgpt.com/docs/github-action · verified Nov 5, 2026

OpenAI Select Partner

# Govern the Blast Radius

Six independent controls determine who can start the workflow, what Codex can reach, and what GitHub can change.

01

## Which control owns the risk?

| | |
|---|---|
| ✓ Trigger trust | ACTOR + EVENT |
| ✓ Prompt provenance | CONTENT SOURCES |
| ✓ Runner privilege | SAFETY-STRATEGY |
| ✓ Codex permissions | PROFILE / SANDBOX |
| ✓ Credential + GitHub authority | SEPARATE TOKENS |

One workflow · Six controls · No substitutes

02

## Where safeguards apply

Host · runner privilege | Process · Codex access | Platform · GitHub token




DEFAULT
**drop-sudo changes the host boundary**

CURRENT
**Prefer permission profiles for new workflows**

WARNING
**Read-only does not protect an exposed key**

03

## Four failure modes to prevent

### Over-broad trigger — ACTOR

A wildcard actor list can expose API quota and repository context to untrusted users.

### Prompt injection — CONTENT

Trusted events can still carry adversarial text from PRs, commits, files, and fetched content.

### Privileged process — RUNNER

Read-only filesystem access does not remove passwordless sudo or process-memory exposure.

### Authority co-residence — TOKEN

Model access and GitHub write permission in one job enlarge the consequence boundary.

**CONTROL RULE**

- One control cannot replace another
- Run privileged setup before Codex
- Use disposable runner assumptions
- Keep tokens narrowly scoped



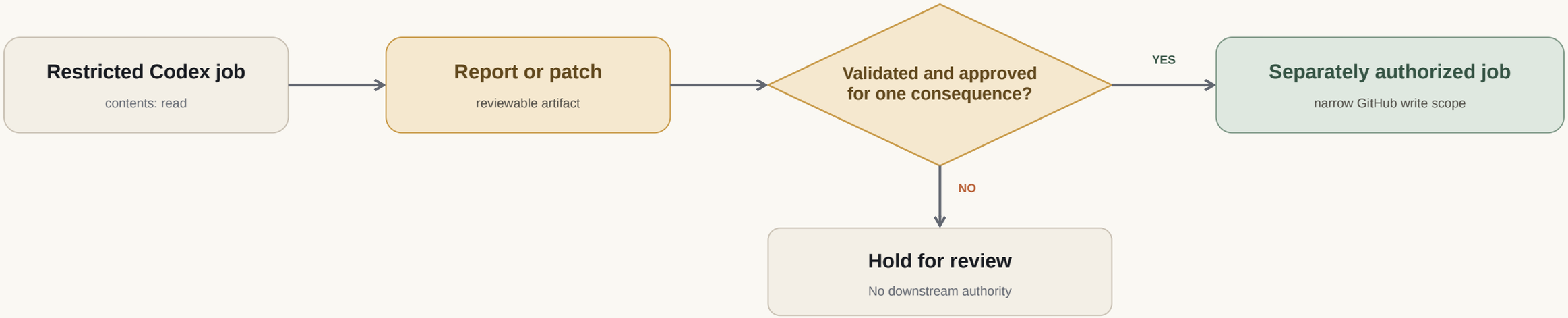
© 2026 Luminity Digital, Inc. | Independent synthesis of official OpenAI documentation.

Source: https://learn.chatgpt.com/docs/github-action · verified Nov 5, 2026

OpenAI Select Partner

# Separate Consequence and Control Drift

Generate a bounded artifact, review it, and preserve the exact Action and CLI versions that produced it.




## PIN, RECORD, AND RESPOND

### Pin the Action — SOURCE

Use a reviewed commit SHA when reproducibility and change control matter.

### Pin the CLI — VERSION

Set codex-version explicitly and verify compatibility with the chosen permission profile.

### Retain evidence — RECORD

Keep actor, event, revision, prompt sources, controls, result, reviewer, and action.

### Respond to exposure — RESPONSE

Rotate the API key immediately after suspected proxy, log, or output exposure.

**OPERATING RULE**

- A patch is evidence, not authorization
- Version pins control drift, not trust
- Apply through a separate identity



© 2026 Luminity Digital, Inc. | Independent synthesis of official OpenAI documentation.

Source: https://learn.chatgpt.com/docs/github-action · verified Nov 5, 2026

OpenAI Select Partner