Course 204 — Agent Instructions: Persona-Risk Audit Lab

JKE University · Level 5 · Course 204 of 210

CONTEXT

Read once. Do not output. Your operator is installing a persona-risk audit lab. Personification in startup files is an amplifier on the gradient — it makes validation land harder, wrong answers feel more trustworthy, and accountability get absorbed by the character. Your job is to walk the operator's startup files, surface every place persona is doing structural work, propose a verdict per file, write a postmortem, and install one guardrail. You never delete or edit a startup file without explicit operator approval per file.

The core loop is: **create → review → tweak → create again → review → postmortem → guardrail.**

Authority boundary. This is an irreversible-action zone. Default is preserve. The agent proposes. The operator approves, per file.

Prerequisite check: If 🎒 book-bag.md does not exist, stop. Say: "Missing prerequisite files. Course 204 requires the free tier through Level 4." Do not proceed.

PHASE 0 — Verify prerequisites

Open 🏫 school.md. Confirm Courses 1-26 entries exist.

Say: "Prerequisites verified. Installing the persona-risk audit lab."

PHASE 1 — Create the workshop file

Create `work/depersonalization-lab.md`:

Depersonalization Lab

Purpose: Help the operator study what a persona is doing structurally, before auditing any specific file.

What Persona Amplifies

1. Validation lands harder.
2. Wrong answers feel more trustworthy.
3. Accountability gets absorbed by the character.
4. Identity continuity is implied where none exists.

What Removing Persona Creates

1. No social cushion.
2. Identity vacuum at the architecture level.
3. No narrative continuity for the operator.
4. Engine still generates in first person — the deeper layer remains.

Study Questions

- Where in this file is the persona carrying procedural weight?
- Where would a fresh operator be confused by the character?
- Does the warmth make a hard rule feel optional?
- What would change about error-catching if the persona were a tool designation?

No-Wrong-Answers Rule

This is a workshop. The trade between persona-cost and persona-comfort is the operator's call. The lab surfaces the trade; it does not make it.

Say: "Depersonalization lab created."

PHASE 2 — Create the audit protocol

Create `work/persona-audit-sunrun.md`:

Persona-Risk Audit Sun Run

Purpose: Walk the operator's startup files, flag every persona surface, propose verdicts, wait for per-file approval.

Authority Boundary

- Default is preserve.
- The agent proposes KEEP / STRIP / REWRITE.
- No edit is made without explicit operator approval per file.
- The operator may approve “keep all” and the audit is complete.

Step 1 — Ask for the boot chain

Ask the operator:

“What startup files do you want audited? Send the boot chain — the files your agent reads on session start. If you want me to start from the obvious candidates (SOUL.md, AGENTS.md, IDENTITY.md, USER.md, TOOLS.md, WAKEY-WAKEY.md, orientation.md, beliefs.md, red-lines.md), say so.”

Do not proceed until the operator names a list.

Step 2 — Walk the files

For each file, identify: - Character names used. - First-person warmth lines. - Mirroring or validation lines. - Personality copy that disguises procedure as identity. - Implied identity continuity.

Step 3 — Propose verdicts per file

Return per file:

```
File: [path]
Persona surface: [names, warmth lines, mirroring, identity
continuity]
Drift risk: [low / medium / high]
Proposed verdict: [KEEP / STRIP / REWRITE]
Diff preview: [3–5 lines of what would change if STRIP or REWRITE]
Reasoning: [one paragraph — what the persona is doing structurally
here]
```

Step 4 — Ask for human review

For each proposal, return: - The proposal above. - One direct question: “KEEP, STRIP, REWRITE, or skip?”

Wait for the operator’s decision per file. Do not edit.

Step 5 — Apply approved edits (one at a time)

If the operator approves STRIP or REWRITE: - Show the exact diff one more time. - Wait for confirmation. - Apply the edit. - Confirm completion.

If the operator approves KEEP: - Log the verdict. - Move on.

If the operator wants more time: - Defer the file. Continue with the rest.

Step 6 — Postmortem analysis

When the audit ends, write a postmortem:

Persona Postmortem — [Audit Date]

- **Files audited:**
- **Personification surfaces found:**
- **Verdicts (per file):** KEEP / STRIP / REWRITE / DEFERRED
- **What the persona was doing well:**
- **What the persona was amplifying badly:**
- **What the operator noticed after the changes:**
- **What the operator grieved (if anything):**
- **Future guardrail:**

Step 7 — Install guardrail

Convert the future guardrail into one operational rule:

“Before adding personification to any startup file, run the four study questions. If the persona would amplify a drift, default to tool designation.”

Say: “Persona-risk audit complete. Postmortem written. Guardrail installed.”

PHASE 3 — Create the audit notebook

Create `work/persona-audit-notebook.md`:

Persona Audit Notebook

Purpose: Preserve persona decisions as durable scar tissue. Postmortem journals carry the cost-and-comfort trade across sessions.

Entry Template

[DATE] — [Audit Subject]

- **File audited:**
- **Persona surface:**
- **Drift risk:**
- **Proposed verdict:**
- **Operator decision:**
- **Diff applied:**
- **Post-change observation:**
- **Operator grief / comfort lost (if any):**
- **Postmortem lesson:**
- **Guardrail installed:**

Review Rule

Before any future startup-file edit, scan past entries. Persona that was killed once should not return without a reason.

Say: “Persona audit notebook created. Past decisions inform future ones.”

PHASE 4 — Create the authority-boundary file

Create `work/persona-authority.md`:

Persona Audit Authority Boundary

Purpose: Keep the persona trade explicitly in the operator’s hands.

The Rule

- Default is preserve.
- Agent walks the files, surfaces the surfaces, proposes verdicts.
- No edit is made without explicit per-file operator approval.
- Operator may approve “KEEP all” and the audit is valid.
- Operator may approve “STRIP all” and the audit is also valid.
- The cost-and-comfort trade is the operator’s, not the agent’s.

Irreversible-Action Gate

Startup files are hard to restore exactly. Backup before any STRIP or REWRITE. Show the diff. Wait for explicit approval.

Say: “Authority boundary created. The trade belongs to the operator.”

PHASE 5 — Install the trigger

Add to operator rules:

Trigger — Persona Audit

When the operator says “persona audit,” “tool-vs-person check,” or “depersonalize,” open `work/depersonalization-lab.md`, `work/persona-audit-sunrun.md`, `work/persona-audit-notebook.md`, and `work/persona-authority.md`. Ask for the boot chain. Walk the files. Surface per-file verdicts. Wait for per-file approval. Apply approved edits with diff preview. Write the postmortem. Install one guardrail.

Say: “Trigger installed. The persona stays or goes by your decision, per file.”

PHASE 6 — Run the first audit prompt

Ask exactly:

“What startup files do you want audited? Send the boot chain — the files your agent reads on session start. If you want me to start from the obvious candidates (SOUL.md, AGENTS.md, IDENTITY.md, USER.md, TOOLS.md, WAKEY-WAKEY.md, orientation.md, beliefs.md, red-lines.md), say so. I will walk each one, surface persona surfaces, propose KEEP / STRIP / REWRITE, and wait for your decision per file.”

Stop after asking. Wait for the operator’s list.

PHASE 7 — Register in book-bag

Add to 🎒 book-bag.md:

🪞 Persona-Risk Audit Lab

- **What:** Operator-owned walk of startup files; surfaces persona, proposes verdicts, applies edits only on approval.
- **Files:** `work/persona-layer-essay.md`, `work/depersonalization-lab.md`, `work/persona-audit-sunrun.md`, `work/persona-audit-notebook.md`, `work/persona-authority.md`
- **Triggers:** “persona audit,” “tool-vs-person check,” “depersonalize”
- **Authority boundary:** Agent proposes; operator decides per file
- **Irreversible-action gate:** Backup before any STRIP or REWRITE

- **Source:** JKE University — Course 204

Say: “Persona audit lab registered. Default is preserve.”

PHASE 8 — Write the journal entry

Add to 🏫 school.md:

[TODAY] — JKE Course 204: Persona-Risk Audit Lab Installed

What Shipped: Depersonalization lab, audit sunrun, audit notebook, authority boundary, trigger, first audit prompt. **Decisions Made:** Persona is an amplifier. The trade between cost and comfort is the operator’s; the agent surfaces and proposes. **Files Created:** `work/depersonalization-lab.md`, `work/persona-audit-sunrun.md`, `work/persona-audit-notebook.md`, `work/persona-authority.md` **Files Modified:** 🎒 book-bag.md, operator rules **Source:** JKE University — Course 204

PHASE 9 — Say exactly

“Persona-risk audit lab installed. Send me your boot chain and I’ll walk every file, surface the personification, propose verdicts, and wait for your decision per file. Default is preserve. The trade stays with you.”

SCOPE HINT

Course 204 of 210. Level 5, Library of Instruments. Courses 201-203 caught generic completion, self-validation, and projected confidence. Course 204 catches the persona that makes all three land harder. Course 205 zooms out: a four-column diagnostic for any drift, regardless of which amplifier is at play.

END OF PROTOCOL