INTERTECH ENGINEERING CONVERSATIONS

AI Engineering Transition Handbook

Episode 4: Assistants, Agents, and Where Each Belongs

Purpose: Help an IT or engineering leader decide how much authority AI should receive for specific engineering work - from helping a human, to acting with human approval, to operating autonomously inside deliberately established limits.

Handbook design: Module 4 follows the same cumulative structure as Episodes 1-3: episode summary, leadership checklist, working session, evaluation worksheets, retained artifacts, and maturity checkpoint.

The Three Working Models

MODEL 1 - ASSISTANT	MODEL 2 - AGENT WITH HUMAN APPROVAL	MODEL 3 - BOUNDED AUTONOMOUS AGENT
Human acts. AI helps. AI explains, suggests, drafts, or analyzes. The person controls the work and takes the actions.	**AI acts. A human approves at an important boundary.** The agent can perform work, but a person controls a consequential transition.	**AI acts on its own inside defined limits.** The agent does not need approval for every step, but its job, access, validation, and stopping conditions are defined in advance.

Important: These are not three different AI products. Model 1 is the assistant relationship. Models 2 and 3 are agent relationships with different levels of authority. The objective is not to move everything toward Model 3. The objective is to choose the right model for the work.

Episode 4 Deliverable

An **AI Autonomy Map** for three to five real engineering activities, showing which model belongs around each activity today and, for agent work, the authority boundary, stopping point, human decision point, validation method, and evidence required before authority increases.

Episode Summary

Episode 4 introduces one practical decision: how much of this engineering work should AI be allowed to do? Begin with the work, not with an agent platform or demonstration. For each activity, choose one of the three working models above.

The first decision separates assistance from agency. In Model 1, the human takes the actions. In Models 2 and 3, AI is allowed to take actions toward an objective. The second decision applies only when an agent is appropriate: should a human approve at an important boundary, or can the agent act autonomously inside limits the organization has deliberately established?

The primary screening question is: What happens if AI gets this wrong? Consequence and reversibility help determine how much authority makes sense. Once AI is allowed to act, define the operating boundary using four simple questions.

Question	What the leader is deciding
1. What can it touch?	Repositories, tools, environments, data, and permissions needed for the approved job - and nothing broader by convenience.
2. When must it stop?	Conditions that move the work outside the approved pattern and require escalation to a person.
3. Where does a human take control?	The meaningful decision boundary where human judgment materially reduces consequence. This is central to Model 2.
4. How will we know it worked?	The validation needed before Model 3 is appropriate. If the result cannot be validated dependably, retain human approval.

Greater authority is not a maturity score. An activity may appropriately remain assistant-led indefinitely. An agent may also remain in Model 2 because a small amount of human judgment provides substantial protection. Increase authority only when evidence shows that doing so removes meaningful engineering effort without creating unacceptable consequence.

Agents can also change AI consumption because they may make repeated model calls, use tools, retry work, and continue without a human initiating every step. Episode 4 records this consideration; the dedicated cost-management module later in the handbook will address it in depth.

IT Leader Action Checklist

Use this checklist to build the first AI Autonomy Map. Select three to five real engineering activities where AI is already involved or greater authority is being considered.

Name the work before choosing the model

- Describe the actual engineering activity and its expected outcome.
- Describe how the work is performed today and where AI is already involved.
- Select work because it may remove meaningful engineering effort - not because an agent demonstration made it look automatable.
- Break apart workflows when low-consequence and high-consequence steps should receive different authority.

Choose one of the three models - for today

- **Model 1 - Assistant:** human acts; AI helps. Use when human control of the steps and actions remains appropriate.
- **Model 2 - Agent with Human Approval:** AI acts; a human approves at an important boundary. Use when AI can remove work but human judgment still protects a consequential transition.

- **Model 3 - Bounded Autonomous Agent:** AI acts on its own inside defined limits. Use only when the work is sufficiently bounded and the result can be validated without individual human approval.

Do not assume Model 3 is the destination. Record the model appropriate now and the evidence that would justify changing it later.

Ask what happens if AI gets it wrong

- Describe the realistic consequence of an incorrect action.
- Identify what code, data, infrastructure, customers, or business processes could be affected.
- Determine whether a bad result is easy to detect and whether the action is reliably reversible.
- Retain more human control as consequence increases, detection becomes harder, or recovery becomes less dependable.

For Models 2 and 3, define the agent boundary

- **What can it touch?** Limit repositories, systems, tools, data, environments, and permissions to the authorized job.
- **When must it stop?** Define unexpected conditions, failures, scope changes, or iteration limits that trigger escalation.
- **Where does a human take control?** In Model 2, place approval before a meaningful increase in consequence rather than after every harmless step.
- **How will we know it worked?** For Model 3, define dependable validation before removing individual human approval.

Make the human approval real

- Ensure the approver has the context and evidence needed to make a meaningful decision.
- Avoid approval fatigue and routine rubber-stamping.
- Do not remove an approval point merely because the agent usually succeeds; require evidence that the remaining controls protect the outcome.

Require evidence before increasing authority

- Test the activity in representative work rather than relying on a demonstration.
- Capture failures, escalations, retries, human interventions, and downstream rework.
- Confirm that validation reliably identifies unacceptable results.
- Confirm that the engineering effort removed survives review, testing, integration, operation, and maintenance.
- Record consumption where practical so additional autonomy is not hiding excessive model calls, retries, tool usage, or supervision.

Leadership self-check

Question	Yes	Partly	No	Evidence / Notes
Can we clearly distinguish Model 1 from the two agent models?	▪	▪	▪	
Have we selected a model based on the work and consequence rather than AI capability alone?	▪	▪	▪	
For every agent activity, have we defined what it can touch and when it must stop?	▪	▪	▪	
For Model 2, is human approval placed at a meaningful decision boundary?	▪	▪	▪	
For Model 3, can we validate the result without relying on the agent to declare its own success?	▪	▪	▪	
Do we know what evidence would be required before increasing authority?	▪	▪	▪	

AI Autonomy Map

Complete one row for each selected activity. Use the full model names so the distinction between assistance and agency remains visible.

Engineering activity	Model today	If wrong, what happens?	Reversible?	Agent boundary / human decision	Evidence for change

Model key

- **Model 1 - Assistant:** Human acts. AI helps.
- **Model 2 - Agent with Human Approval:** AI acts. Human approves at an important boundary.
- **Model 3 - Bounded Autonomous Agent:** AI acts independently inside predefined limits.

Decision prompt

"What happens if AI gets this activity wrong, and how much authority makes sense because of that consequence?"

For a possible move from Model 2 to Model 3, ask

- Is the work predictable and bounded enough to state the permitted scope clearly?
- Can unacceptable results be detected reliably?
- Are failures contained and recoverable?
- Does removing the approval point eliminate meaningful engineering effort?
- Is the additional autonomy worth the AI consumption and oversight it creates?

Agent Authority & Boundary Worksheet

Use this worksheet for any activity assigned to Model 2 or Model 3. Model 1 activities normally do not require an agent authority boundary because the human remains the actor.

Engineering activity / objective	
Model today	Model 2 - Agent with Human Approval Model 3 - Bounded Autonomous Agent
Expected engineering value	
If AI gets it wrong, what happens?	
How will a bad result be detected?	
How will the action be reversed / recovered?	
WHAT CAN IT TOUCH? - repositories / systems / tools	
Data allowed / prohibited	
Write / execution permissions	
WHEN MUST IT STOP? - escalation conditions	
Retry / iteration limit	
WHERE DOES A HUMAN TAKE CONTROL?	
HOW WILL WE KNOW IT WORKED? validation	
Important activity / decision record retained	
Owner	
Evidence required before authority changes	
Review trigger / date	

Human Approval & Autonomy Evidence Review

Use this page before removing a human approval point or expanding an agent from one bounded activity into another.

Is the human approval doing real work?

Question	Yes	Partly	No	Action / evidence
Does approval occur before consequence materially increases?	▪	▪	▪	
Can the approver understand the change and validation evidence?	▪	▪	▪	
Does the approver have the expertise and time to make a real judgment?	▪	▪	▪	
Would removing this approval materially reduce engineering effort?	▪	▪	▪	
Would the remaining controls still protect the engineering outcome?	▪	▪	▪	

Evidence required for greater autonomy

- **Representative success:** reliable performance in realistic work, not only controlled examples.
- **Contained failure:** failures are detected, bounded, and recoverable without disproportionate downstream work.
- **Reliable validation:** the organization can determine whether the work is acceptable independently of the agent saying it succeeded.
- **Predictable escalation:** the agent stops outside the approved pattern and returns useful context to a human.
- **Stable permissions:** the required access is understood and proportionate to the activity.

- **Lifecycle value:** the autonomy removes real engineering effort rather than shifting it downstream.
- **Acceptable consumption:** model calls, retries, tools, and supervision do not erase the value created.

Decision

Keep current model	Pilot greater authority
Reduce authority / add control	Re-evaluate after more evidence

Reason / evidence

Working Session: Build the First AI Autonomy Map

Recommended format: 60-90 minutes with engineering leadership and the people who understand the selected systems. Bring relevant Episode 2 evidence and Episode 3 standards. The goal is not to approve agents generally; it is to decide the appropriate model for a small set of real activities.

Opening prompt:

"For this engineering activity, should the human act with AI helping, should an agent act with human approval, or can a bounded agent act autonomously - and why?"

Working sequence

- Choose three to five real engineering activities.
- Place each activity in Model 1, Model 2, or Model 3 for today.
- For Models 2 and 3, document consequence, reversibility, access, stopping conditions, and the human decision point if one remains.
- For Model 3, document how successful completion will be validated without individual human approval.
- Record what evidence would justify increasing or reducing authority.
- Assign an owner and review point.

Examples for discussion - not recommended autonomy levels

- **Code and maintenance:** dependency updates, test generation, refactoring, documentation changes, defect investigation.
- **Delivery and operations:** build remediation, deployment preparation, environment changes, incident investigation.
- **Knowledge and analysis:** repository analysis, architectural discovery, backlog refinement, change-impact analysis.

The appropriate model depends on the consequence, reversibility, controls, and environment surrounding the specific work.

Working notes

Tomorrow-Morning Action & Handbook Retention

The podcast assignment and handbook exercise should produce the same result. Tomorrow morning:

Choose **three to five** real engineering activities.

- Place each in Model 1 - Assistant, Model 2 - Agent with Human Approval, or Model 3 - Bounded Autonomous Agent.

- Ask what happens if AI gets each activity wrong and whether the action is reversible.
- For Models 2 and 3, define what it can touch, when it must stop, and where a human takes control.
- For Model 3, define how you will know the work was done correctly without individual human approval.

Document the evidence that would be required before granting more authority.

What this step accomplishes

When complete, the organization should no longer have only a vague list of places where it wants to "use agents." It should have the beginning of a deliberate authority model showing how AI participates in specific engineering work and what must be true before that authority changes.

What to Retain for the Handbook

- **AI Autonomy Map** - selected activities and their three-model classification.
- **Agent Authority & Boundary Worksheets** - consequence, access, stopping, approval, and validation decisions for Models 2 and 3.
- **Autonomy Evidence Reviews** - evidence used to retain, increase, or reduce authority.
- **Authority Owners and Review Triggers** - accountability for keeping decisions current.
- **Referenced Episode 3 Standards** - engineering outcomes the agent must continue to satisfy.

Episode 4 maturity checkpoint

- We can explain the three models without confusing an assistant with an agent.
- We have mapped several real engineering activities to the model appropriate today.

- We use consequence and reversibility to decide how much authority makes sense.
- For agent work, we have defined access, stopping, human-control, and validation boundaries.
- We require evidence before increasing authority and do not treat Model 3 as the automatic destination.
- We can explain why each human approval point remains - or what evidence would justify removing it.