MODULE 2 · SESSION 2

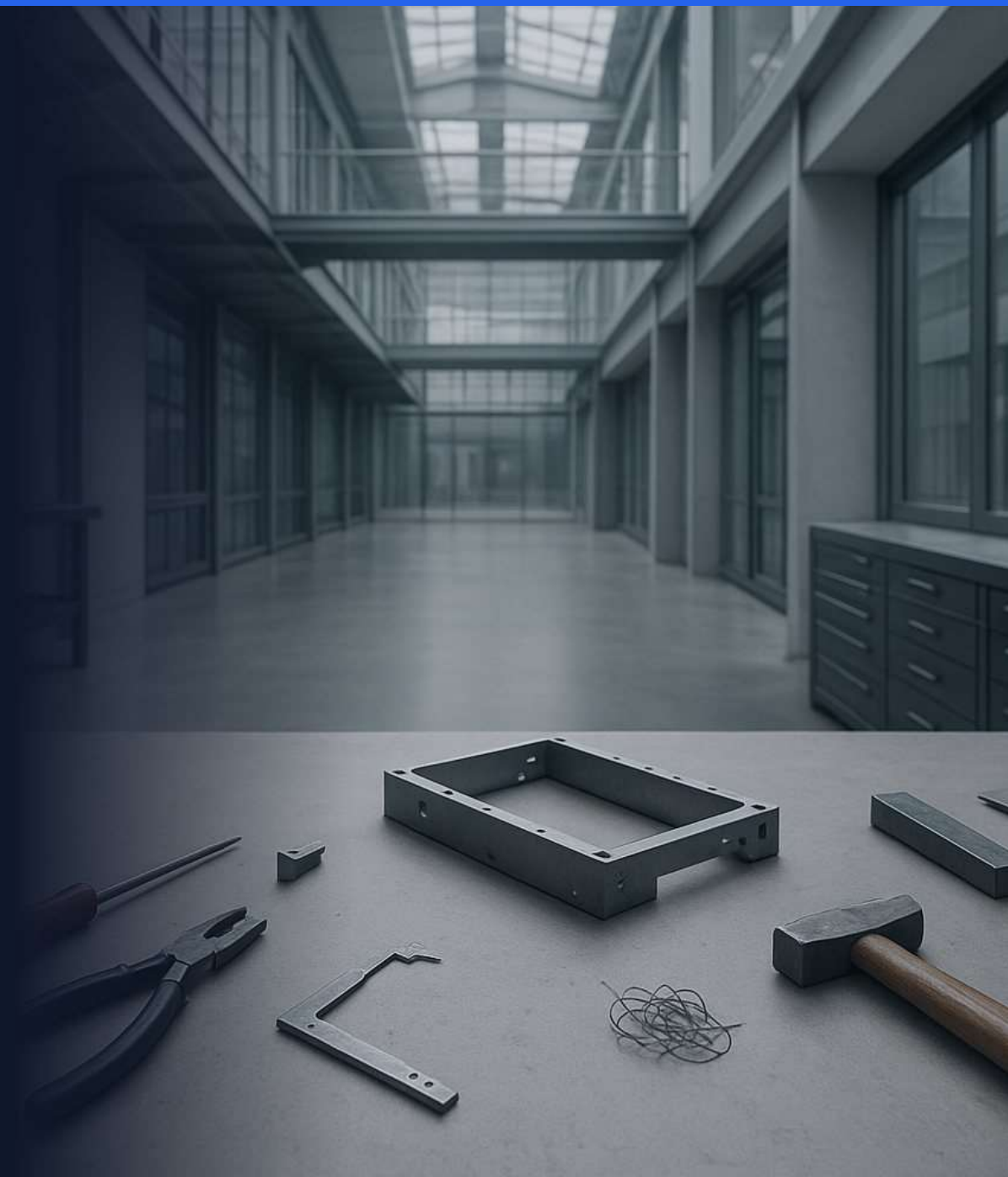
Bound Autonomy With Human Escalation

How to decide which decisions the system may make on its own, and design the path for everything else.

Theory session

Beginner

Running case: HarborPilot Health



ORIENTATION

Where this session sits

2.1 measured the system; 2.2 decides what it may do alone

MODULE 1
AI System Foundations

MODULE 2
Controlled AI Operations

MODULE 3
Strategic AI Operating Choices

SESSIONS IN THIS MODULE

2.1 Measure Reliability Before Scale
An evaluation loop that detects quality drift

2.2 **Bound Autonomy With Human Escalation**
What the system may decide alone, and what it must escalate
YOU ARE HERE

2.3 Defend and Recover AI Services
Detect an attack, contain it, and restore trust

SESSION OUTCOMES

What this session explains

Setting limits that make automation safe and still fast

01

Autonomy is not one switch

It is a level, set separately for each decision the system makes.

02

Two questions set the level

How reversible is the action, and how large is the damage if it is wrong?

03

Escalation must be designed

Who receives it, with what context, in what time, and what happens if nobody answers.

04

Review gates can be cheap

Mayo Clinic added a human gate and still went from monthly to weekly forecasting.

CONCEPT 1 · LEVELS

Autonomy comes in degrees

A simplified version of the classic automation ladder

ACTS ALONE

The system decides and acts; humans see a summary afterwards, if at all.

ACTS UNLESS STOPPED

The system announces what it will do and proceeds after a delay unless someone objects.

RECOMMENDS, HUMAN APPROVES

The system prepares the decision; a person authorises it. Mayo Clinic's review gate.

ASSISTS

The system organises information; the person decides everything.

Sheridan, T. & Verplank, W. (1978). Human and Computer Control of Undersea Teleoperators. Cambridge, MA: MIT Man-Machine Systems Laboratory.

THE KEY IDEA

Set the level per decision

One product normally runs at several levels at once — assembling at the top, approving at the bottom.

THE MAYO PATTERN

High autonomy inside a scope

Agents worked autonomously within a defined sub-domain; consequential decisions still required approval.

THE TRAP

Acts unless stopped

Looks like oversight and usually is not — nobody has time to object to everything.

CONCEPT 2 · THE TEST

Two questions that set the level

Reversibility and blast radius, asked for every action

SIZE OF THE DAMAGE ▲

APPROVE FIRST Large and hard to undo. A human authorises before anything happens.	**ACT, THEN REVIEW** Large but recoverable. Automate with a prompt, visible audit trail.
NOTIFY Small and hard to undo. Automate, and tell someone it happened.	**AUTOMATE FULLY** Small and easily reversed. No human in the path at all.

HOW REVERSIBLE THE ACTION IS →

WHY REVERSIBILITY FIRST

It is cheaper than accuracy

Making an action undoable often costs less than making the model good enough to never be wrong.

THE DESIGN MOVE

Change the quadrant

Add an undo, a hold period or a cap, and a top-left action becomes a top-right one.

CONCEPT 3 · ESCALATION

What a usable escalation path specifies

Six things; most designs stop after the first two

Element	The question	HarborPilot example
Trigger	What causes an escalation?	Confidence below threshold, a value above a cap, or an unfamiliar request type.
Recipient	Who exactly receives it?	The on-shift logistics analyst, by role and not by name.
Context	What do they need to decide quickly?	The request, the system's recommendation, its uncertainty, and the two nearest precedents.
Time limit	How long may it wait?	Fifteen minutes for a delivery decision; one working day for a plan revision.
Default on no answer	What happens if nobody responds?	Hold and repeat — never proceed silently on a high-stakes action.
Feedback	Does the decision come back to the system?	Every human override is recorded as a labelled example.

THE ROW THAT MATTERS MOST

The fifth row is the one that causes real incidents. An escalation with no defined timeout quietly becomes full autonomy on the busiest day of the year.

CONCEPT 4 · TWO WAYS IT GOES WRONG

Trusting too much and trusting too little

Both are well documented, and both are expensive

AUTOMATION BIAS

Accepting the system's answer

WHAT IT LOOKS LIKE

Approvals granted in seconds, without the evidence being examined.

WHY IT HAPPENS

The system is usually right, so checking feels like wasted effort.

THE FIX

Show uncertainty prominently, and require a reason for approval on high-stakes items.

DISUSE

Ignoring the system

WHAT IT LOOKS LIKE

People work around it, or override routinely.

WHY IT HAPPENS

Too many false alarms, or recommendations that do not explain themselves.

THE FIX

Tune thresholds to reduce noise, and show the reasoning, not just the answer.

THE TEST FOR A REAL GATE

A review gate that is always approved is not oversight. Measure the override rate: if it is near zero or near one hundred per cent, the gate is not doing anything.

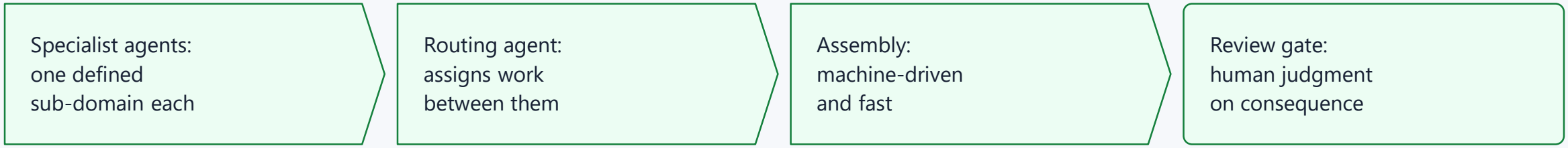
Parasuraman, R. & Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. Human Factors, 39(2), 230–253.

CONCEPT 5 · SCOPE

How Mayo Clinic kept autonomy bounded

Specialisation, routing, assembly, and a gate

THE BOUNDED-AUTONOMY PATTERN



WHAT THE BOUNDARY BOUGHT

Autonomy expanded inside each sub-domain while decision rights stayed explicit at the end. Forecasting moved from monthly to weekly, manual operations fell 42%, and revenue rose 14% — the boundary did not cost speed.

CONCEPT 6 · CALIBRATION

How much to escalate

The rate itself is the thing to watch

TOO LITTLE

Almost nothing escalates. The gate exists on paper and consequential errors pass through it.

CALIBRATED

Escalations are few enough to be examined properly and frequent enough to catch real problems.

TOO MUCH

Everything escalates. Reviewers rubber-stamp, and the gate becomes a formality.

THREE NUMBERS TO WATCH WEEKLY

Metric	What it tells you
Escalation rate.	Whether the threshold is in a workable range at current volume.
Override rate within escalations.	Whether the gate is changing outcomes or just adding delay.
Time to decision.	Whether reviewers actually have the capacity the design assumes.

CONCEPT 7 · FAILURE MODES

Three ways bounded autonomy stops being bounded

All three happen gradually

The timeout that proceeds

Looks like "If no response in ten minutes, continue."

Why it fails On the busiest day, nobody responds and everything auto-approves.

Fix Hold by default on high-stakes actions; escalate further rather than proceeding.

Scope creep by capability

Looks like The agent gets better, so more is handed to it informally.

Why it fails The written boundary and the real one diverge without a decision.

Fix Autonomy level is a recorded setting changed deliberately, not a drift.

A gate nobody can use

Looks like An approval screen with the answer and no reasoning.

Why it fails Reviewers approve because they have no basis to do anything else.

Fix Show uncertainty, the evidence and the nearest precedents.

CASE STUDY

Mayo Clinic

Autonomous forecasting with an explicit human review gate for consequential decisions.

CASE STUDY · MAYO CLINIC

Machine assembly, human judgment

Scoped agents, a routing agent, and a review gate on consequence

THE DESIGN

Scoped agents, routed work

THE STRUCTURE

A multi-agent forecasting and planning system for clinical logistics, with specialist agents each responsible for a defined sub-domain.

THE COORDINATION

A routing agent orchestrated task assignment across those specialists.

THE BOUNDARY

Assembly was machine-driven; judgment stayed with human analysts behind a review gate.

THE RESULT

Faster, and still bounded

CADENCE

Forecasting moved from monthly to weekly.

EFFORT

Manual operations decreased by 42 per cent.

VALUE

A 14 per cent revenue increase associated with improved forecasting and pricing.

READING THE CASE

Autonomy became a design choice about boundaries rather than a single setting applied to everything. The boundary did not make the work slower.

Drawn from the session case pack.

APPLICATION

Bringing the theory to HarborPilot Health

Setting the boundary for each decision

Step	What you decide	HarborPilot example
List the actions	Every distinct thing the system can do.	Reorder supplies, reschedule a delivery, change a forecast, notify a ward.
Score reversibility and damage	Two simple ratings per action.	Reordering: reversible, moderate. Rescheduling a clinical delivery: hard to undo, high.
Assign the autonomy level	From the four-quadrant map.	Reorder under a value cap: automate. Reschedule: approve first.
Design the escalation	All six elements, in writing.	Trigger, on-shift analyst, context pack, 15-minute limit, hold on no answer, override recorded.
Add reversibility where you can	It is usually cheaper than more accuracy.	A one-hour hold window on automatic reorders makes them safely automatable.
Watch three rates weekly	Escalation, override, time to decision.	Reviewed in the same weekly meeting as Session 2.1's metrics.

WHAT COMES NEXT

Session 2.3 asks what happens when something goes wrong deliberately — and how you recover credibly.

SUMMARY

What to carry out of this session

1. **Autonomy is a level, set per decision.**
   One product normally runs at several levels at the same time.

2. **Reversibility and damage set the level.**
   Adding an undo is often cheaper than making the model good enough never to be wrong.

3. **Design the whole escalation path.**
   Trigger, recipient, context, time limit, default on no answer, and feedback.

4. **Never proceed silently on a timeout.**
   Hold by default, or a high-stakes gate disappears on the busiest day.

5. **A gate always approved is not a gate.**
   Watch the override rate — near zero means it is doing nothing.

FURTHER READING

Sources for this session

Primary references behind each concept

LEVELS OF AUTOMATION	Sheridan, T. & Verplank, W. (1978). *Human and Computer Control of Undersea Teleoperators.* Cambridge, MA: MIT.
TYPES AND LEVELS	Parasuraman, R., Sheridan, T. & Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics*, 30(3), 286–297.
USE AND MISUSE	Parasuraman, R. & Riley, V. (1997). Humans and automation. *Human Factors*, 39(2), 230–253.
AUTOMATION BIAS	Skitka, L., Mosier, K. & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1006.
IRONIES OF AUTOMATION	Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779.
RISK MANAGEMENT	NIST (2023). *AI Risk Management Framework (AI RMF 1.0).* Gaithersburg: NIST.
HUMAN OVERSIGHT	Shneiderman, B. (2022). *Human-Centered AI.* Oxford: Oxford University Press.