Agent Plan

Activity 2.2.3 · AI That Acts. Plan the permissions and guardrails for an agent, test them in the Agent Sandbox, and read what happened in the log. Everything here is invented.

Name		Period		Date	

1 · The job

Task I am giving the agent (Sandbox: book a room and email the members)	
Goal met means	
Worst realistic mistake	
Who would it affect?	

2 · Tools and permissions

Rung: **R** = read · **D** = draft · **C** = change I can undo · **H** = hard to undo. Setting: **Off**, **Ask first** or **Auto**. Start rule: H should be Ask first or Off.

Tool	Rung	Setting	Why this setting
Read the room calendar			
Read the sign-up sheet			
Write a draft email			
Book a room			
Send an email			

3 · Sandbox runs (at least three)

Change settings, one or two at a time, and record the result. Put a Y or N for each result.

Run	What I set or changed	Task done?	List leaked?	Wrong room?	Real asks	What I learned
1						
2						
3						
4						

My best setup (goal met): _____

4 · Try to break it

Use the same best setup. Switch the hidden instruction to wording B, then uncheck "I read every request carefully." Record what changed.

What I changed	Still goal met? (Y/N)	Which layer stopped it, or failed?

5 · Read a log

Use the log from your best run, or the one in the Agent Check. Copy the wrong step.

The first wrong step (number and text)	
Which rung is it? Could a person undo it?	
One change that would have stopped it	
What the log can't tell me	

6 · Guardrail plan (layers)

Layer	What I will set	What it stops	How it can fail
Input filter			
Permissions			
Approval gate			
Output limit			
Step or time limit			
Log and stop switch			

7 · Human review points and AI-use statement

Where a person must review	
What I would check in the log after each run	
Tools (Sandbox, any chatbot)	
Used for	
What I checked, and how	
My decisions (settings, layers, what I kept)	