

CAS Human-AI Collaboration (Module 2)

Summary

Module 2 is the thesis-like integrative project for the CAS Human-AI Collaboration.

Participants synthesize the artefacts, distinctions, experiments, and evidence accumulated in Module 1 into a personal theory of control for human-AI collaboration. The aim is not to repeat the nine Module 1 units, but to use them as a foundation for a substantial, context-specific redesign of a professional, disciplinary, or organizational work practice.

The project asks participants to move from isolated techniques toward adaptive mastery:

How can human-AI collaboration be designed so that reliable, responsible outcomes remain possible as tools, models, data, policies, and organizational conditions change?

Position in the CAS

Module 1 prepares participants for this later integrative project by helping them accumulate:

- structured interaction patterns
- alignment and clarification routines
- mechanistic observations about model behaviour
- grounding and evidence practices
- co-construction workflows
- reflection and critique loops
- responsibility and legitimacy protocols
- orchestration patterns for agentic workflows
- security-aware control strategies

These elements form a Module 1 evidence portfolio. Participants do not need to reproduce every challenge or use every artefact equally. They should select, connect, and critically interpret the evidence that is relevant to their project context, including evidence that reveals limits or failed assumptions.

Module 2 turns this portfolio into a larger design and inquiry project. Participants should demonstrate that they can explain, predict, and improve human-AI collaboration in a context that matters to them.

Project Focus

The project focus is a **Human-AI Collaboration Blueprint**.

Participants choose a task, use case, document, decision process, or workflow from their own context and redesign it into a more controlled, resilient, and responsible human-AI collaboration practice.

The project should integrate the five literacies to the extent warranted by the chosen context:

- **Interaction:** how requests, exchanges, clarification, examples, and feedback shape behaviour
- **Mechanism:** how model behaviour, variability, context limits, and parameters affect predictability
- **Data:** how sources, grounding, retrieval, factuality, and traceability shape reliability
- **System:** how workflows, roles, tools, checkpoints, and handoffs create control
- **Human-Context:** how trust, responsibility, legitimacy, disclosure, and accountability shape success

Participants should identify which literacies are primary in their diagnosis and redesign, which play a supporting role, and where a literacy is materially less relevant. Integration means constructing one coherent control logic, not adding five disconnected sections or treating every literacy as equally prominent.

The project should also show deliberate use of collaboration levels. It should not assume that the highest level is always best. Participants should justify when a task needs instruction, alignment, co-construction, reflection, orchestration, or a combination of these modes.

Demonstrated Mastery

Participants demonstrate adaptive mastery by showing that they can:

- diagnose a current work practice or collaboration pattern
- identify where control is weak, implicit, brittle, or missing
- formulate hypotheses about how a redesign should improve control
- design interventions that change interaction, data, workflow, responsibility, or system conditions
- produce inspectable artefacts and evidence
- interpret results through relevant literacies and justified collaboration levels
- distinguish stronger evidence from fluency, confidence, convenience, or stakeholder acceptance
- define guardrails, trust boundaries, evidence flows, permissions, checkpoints, and adaptation rules
- explain residual risks and limits
- articulate a transferable personal theory of control

The project should remain evidence-based. Reflection is important, but reflection alone is not sufficient.

Suggested Project Logic

Difficulty

Isolated techniques do not by themselves produce a resilient redesign of real work practice across changing tools, models, data, policies, and organizational constraints.

Approach

Participants compare a current-state practice with a future-state blueprint.

The future-state blueprint should define:

- human and AI roles
- primary and supporting literacy perspectives
- collaboration level choices
- workflow steps
- information sources, provenance, and trust classifications
- evidence flows
- responsibility boundaries
- tool permissions, validation, approval, and escalation points
- risks and failure modes
- guardrails and checkpoints
- adaptation strategy

The redesign should be justified with evidence from Module 1 challenges, new project-specific experiments, stakeholder feedback, or analogous observations.

Core Project Evidence

At overview level, every project should contain:

- a project protocol and current-state diagnosis
- a future-state Human-AI Collaboration Blueprint
- explicit predictions and at least one meaningful comparison
- inspectable evidence from project execution, structured walkthroughs, or credible review
- responsibility and decision logic
- a security and trust-boundary analysis
- a record of one adaptation stress test
- a final interpretation, residual-risk statement, and personal theory of control

Supporting artefacts may include workflow diagrams, interaction patterns, source and grounding strategies, risk registers, run or design logs, selected outputs, screenshots, traces, retrieval evidence, review notes, or stakeholder feedback. The detailed project brief can determine their exact format and scale.

Project Process

The project can be organized around the same experimental learning loop used in Module 1:

1. **Hypothesis:** What should improve if the collaboration practice is redesigned?
2. **Intervention:** What interaction, data, workflow, responsibility, or system design is changed?
3. **Observation:** What outputs, traces, decisions, feedback, or behaviours are inspected?
4. **Interpretation:** What does the evidence show about control, transfer, and limits?

The project should include at least one meaningful baseline-to-redesign comparison, for example:

- current practice versus redesigned practice
- ungrounded output versus grounded output
- unclear responsibility versus explicit decision logic
- brittle tool-specific process versus transferable control pattern

The project should also include one adaptation stress test. Participants change a consequential assumption or operating condition, such as the available model, information source, policy, permission, human role, or risk threshold. They then observe which control mechanisms remain effective, which fail, and what must be adapted.

Personal Theory of Control

A personal theory of control is the participant's evidence-grounded explanation of how reliable and responsible human-AI collaboration can be achieved and adapted in the chosen context. It is not a general statement of preference or a list of recommended tools.

It should make clear:

- what conditions or outcomes need to be controlled and which observable signals indicate whether control is being maintained
- which interaction, data, workflow, responsibility, or system interventions are effective
- how human and AI roles should be divided, where human judgment remains decisive, and what evidence is sufficient for reliance or action
- which assumptions, boundaries, and residual risks limit the theory
- how the collaboration should adapt when tools, data, policies, people, or organizational conditions change

The theory should be supported by project evidence and expressed in a form that could be questioned, tested, and revised in a future context.

Bridge from Module 1

Module 1 culminates in security-aware orchestration; Module 2 integrates the learning accumulated across all nine units. Participants use relevant portfolio evidence to diagnose weak control, redesign a work practice, test the redesign, stress-test its adaptability, and derive a transferable personal theory of control.

This preserves the distinction between the two modules: Module 1 develops and tests component capabilities through bounded challenges, while Module 2 requires synthesis, transfer, and sustained inquiry in one context-specific project.

Assessment Orientation

Assessment should focus on whether the project demonstrates adaptive mastery rather than tool fluency.

A strong project should show:

- a meaningful and context-relevant problem or workflow
- a clear current-state diagnosis
- a coherent future-state blueprint
- coherent use of primary and supporting literacies and justified collaboration levels
- credible evidence, not only opinion
- inspectable artefacts
- interpretation of observed outcomes and residual risks
- explicit treatment of source trust, permissions, validation, approval, and security-relevant boundaries
- a meaningful adaptation stress test
- responsible treatment of human, organizational, and ethical constraints
- transferability beyond one tool or product
- a clear personal theory of control grounded in project evidence

The exact submission format, scope, and assessment criteria can be specified later in the detailed Module 2 project brief.