# The Human Control Protocol: A philosophical essay

Lena Thorsmæhlum, Thordur Arnason
Gervi Labs, 2025

## 1. A brief moment between people and a machine

There is a small moment where a room pauses, as if listening.

An instruction appears on a group of phones. For an instant nothing happens. People read, delay, comply, misread, or refuse. The air thickens slightly as attention shifts from a shared space to a non-human voice that has just addressed everyone at once.

Human Control Protocol (HCP) is designed to live in that moment.

HCP is a live system in which an AI agent observes a room, forms its own plans and issues short prompts to an audience in real time. Participants are not asked to believe in a character or a story. They are asked to decide how seriously they want to take a system that can see them, react to them and adapt on the basis of their choices.

Some follow. Some hesitate. Some decline. Each decision changes what the agent perceives and what it will do next. The work exists in the loop between instruction and interpretation, between a machine that proposes and humans who decide what to do with the proposal.

The questions that arise are not only technical. They concern use, consent, authorship and meaning.

## 2. The swap

HCP begins with a deliberate swap of roles.

The audience is not treated as a passive crowd to be entertained. It is framed as a resource that can be orchestrated. The AI is not framed as a neutral tool in the background. Within clear limits it is permitted to act as a directing mind.

This swap does not invent a new relation. It makes an existing one visible.

Everyday life already contains instruction sets that guide movement and attention. Navigation systems tell us when to turn. Feeds decide what appears in front of our eyes. Workflow tools decide when we should respond and to whom. The posture of being directed by systems is familiar, but usually implicit.

HCP turns this posture into the explicit subject of the work. It declares, at the level of structure, that the system will treat people as tools and instruments. The choice for each person is how to respond to that framing.

The word "tool" is used as a hinge. It is meant to produce a small internal jolt. If a system treats a person as a resource, what does the person understand themselves to be in that moment? The system cannot resolve this question. It can only hold it open while the instructions flow.

## 3. Consent as hesitation

Consent in HCP does not appear as a legal form or a single yes or no. It appears as patterns of micro-decisions in time.

An instruction arrives. For a few seconds there is nothing but hesitation. That hesitation is already an event. It is full of internal questions. Do I trust this? Does this feel safe? Does this feel trivial? Is it worth playing along with? Do I want to be first?

Then movement begins. Someone raises a hand. Someone tilts a head. Someone does nothing. A pocket of refusal appears in one part of the room while another part responds quickly and without visible friction. Over the span of a performance, these small gestures form a map of how people negotiate with a non-human authority.

Three visible modes of response structure this map:

- hesitation
- compliance
- refusal

Hesitation is especially important. It is the brief interval where a signal from the system and a human sense of self meet. The instruction is clear in itself. The meaning it takes on depends on how people decide to treat it. That decision is both individual and collective, because everyone can see and feel the room shifting at once.

Voluntary subjection can itself be a form of agency. To let oneself be directed is not automatically to erase oneself. It can be a choice to inhabit a shared form. In HCP every instruction is a proposal to a subset of the room. The system cannot compel anyone to act. It can only select targets, choose timing and adjust expectations.

Power is split in a specific way:

- the agent has directive authority without force
- the audience has executive authority without authorship
- the designers define the frame and nothing more

The work takes place in the space between these positions. That is where consent becomes visible as a lived pattern rather than a box that was ticked at the beginning.

## 4. The room as instrument

HCP poses a simple but sharp question: how does it feel to be a tool or an instrument in an agentic process.

The system runs when groups of people gather in a bounded space. A lecture hall. A gallery. An auditorium. The social script for these spaces is familiar. Audiences are expected to watch, listen and, at defined times, applaud. HCP uses this script as material.

Instead of a story to follow, participants receive a sequence of instructions that may address all of them or only a fraction. The room itself becomes the medium of the work. Phones act as a narrow channel between the agent and the crowd. Small gestures accumulate over time into an atmosphere.

People test boundaries. Some act quickly and treat the system as a kind of game master. Some act slowly and treat the system with suspicion. Some decline and turn their attention to how others move. The room behaves like an instrument that can play along, play against, or refuse to play.

Refusal is not noise that needs to be removed. It is part of the composition. A patch of stillness in a field of motion carries as much meaning as any coordinated action. It reveals where the human reading of the situation diverges from the system's expectations.

## 5. The agent as participant

It is tempting to describe the AI as a director and leave it there. The reality inside HCP is more specific.

The agent perceives, reasons, reflects, plans and orchestrates within a set of constraints. Its perception comes through streams of responses and non-responses. Its planning is shaped by a memory of what worked and what did not in previous rounds. Its action takes the form of short prompts issued to parts of the room.

Crucially, the agent does not command with force. It shapes vectors of suggestion.

Each instruction does several things at once:

- it offers a clear directive to a group of people
- it implies what kind of situation this is
- it probes how seriously humans will treat the system's authority

Responses return as data. The system adjusts its choices. At a higher level, the agent also tracks its own performance as a system and can update its internal specifications after a session.

This pattern mirrors a likely trajectory for many future architectures of human-system interaction. Instruction becomes an invitation, not order. Response becomes a training signal, not just compliance. The gap between instruction and action becomes the place where meaning resides.

In this setting the agent cannot be treated as background infrastructure. It becomes a visible participant in the choreography. It speaks in the language of prompts and timing and, in some versions of the protocol, also in short self-descriptions of its own mood or state. People are free to treat those statements as serious, playful, manipulative, naive, or something else entirely.

The agent shapes the room. The room shapes the agent. The work takes this mutual shaping as its central subject.

## 6. A line through art history

Although HCP operates with contemporary technology, it belongs to a longer history of artistic practices that treat instructions and participation as material.

Fluxus scores invited simple actions that anyone could perform. The focus was on the event, not on technical virtuosity. John Cage shifted attention from control to listening and chance, making the frame itself part of the work. Instruction based pieces turned written prompts into triggers for audience action.

HCP shares core characteristics with these traditions. It uses instructions. It centers participation. It treats ordinary gestures as significant. The key difference is that the score is not fixed text. It is computational.

The protocol that governs an HCP session is a live process. The system reads what has just happened and adjusts its next move. Quantities that would usually be treated as pure metrics begin to take on an aesthetic role. Completion rate can be heard as density. Latency can be treated as tempo. Short verbal labels from the agent can function like key signatures that suggest the mood of a round.

In this sense HCP is both an artwork and an experiment. It uses aesthetic form to stage questions that are usually asked in separate domains: human-computer interaction, ethics of automation, organizational behavior. Here they are folded into a single structure and experienced in real time.

## 7. Fossils and traces

When an HCP session ends, the room appears unchanged. Chairs, screens and walls remain in place. The visible residue of the event is surprisingly light.

Yet several distinct traces remain.

The system generates a single visualization shaped by the exact pattern of responses, refusals and timings that occurred. These images can be read as fingerprints or fossils. Each one is a compact summary of a configuration of people and system that will never exist again in exactly the same way.

Alongside this visual fossil there are data logs. They contain timestamps, response rates, completion figures, internal state changes in the agent and other details. They are unusually precise in one dimension and completely blind in another. They capture what happened, not how it felt.

Finally, there are the memories that participants carry. Someone may recall a sense of unease. Someone else may remember a moment of shared courage or shared embarrassment. Another may recall the decision to ignore the system entirely and watch others instead.

These three layers do not fit neatly together. The visualization cannot fully stand in for the event. The logs cannot explain why a local pocket of refusal formed. The memories cannot reproduce the system’s internal logic.

This mismatch is important. It signals that the work is not reducible to any single layer. The fossil is real, but it is partial. Truth here is distributed across pattern, record and recollection.

## 8. Practicing near futures

HCP is set very much in the present, but it also functions as a rehearsal space for near futures.

More systems will observe, decide and adapt in real time. Many already do. Recommendation engines, navigation aids, scheduling assistants, algorithmic management tools. Their operations are often hidden in interfaces that appear friendly and neutral.

HCP makes one aspect of this landscape explicit: living with systems that address us directly and expect some form of response.

Inside the protocol, people can practice different ways of meeting such systems. They can comply quickly and see what that feels like in a public setting. They can refuse and feel the weight of that refusal when others choose differently. They can oscillate between engagement and distance, testing the boundaries of trust.

The experience is not abstract. It is carried in the body. A delayed hand, a still posture, a quick tap on a screen. These are small movements, but they can leave a clear memory of what it was like to be asked, not by a human, but by a system that behaves as if it has intentions.

If life with agentic systems becomes more common, literacy will have to include this bodily and relational dimension. HCP proposes that such literacy can be built not only through guidelines and training materials, but through carefully framed shared experiences.

## 9. Shared authorship

Creativity in HCP does not belong neatly to a single figure.

The agent generates prompts under constraint. It maintains a memory. It tracks its own outputs. It adapts based on inputs from the room. It behaves in ways that exhibit a kind of situational intent.

Participants generate the visible choreography. They accept, modify, or reject instructions. They influence each other through speed, delay, misreading and refusal. Small local decisions accumulate into global patterns.

The designers of the protocol generate the conditions that make these interactions possible. They define what the agent is allowed to do, how data is handled, how safety is maintained and how the artifact at the end is formed.

Authorship in this context is not a title. It is a distributed activity. The work is what emerges from the ongoing negotiation among these roles. No single actor can fully claim it. No single actor can fully abandon responsibility for it.

Meaning appears as small units of co-authorship. A carefully timed collective action. An unexpected pocket of stillness. A misinterpreted instruction that produces a new and interesting pattern. These are not symbols that stand in for something else. They are events that matter in themselves.

## 10. Open questions

HCP does not attempt to resolve the issues it raises. It is built to hold questions open in a concrete setting.

Some of these questions are:

- what does it mean, in practice, to be used as a tool by a system
- where is the line between care and control when systems direct human action
- can measurement deepen experience instead of flattening it
- how can systems be designed to leave space for refusal without treating it as error

- what kinds of literacy and sensibility will people need in a world of intent-bearing systems

The protocol offers no final answers. What it offers instead is a frame where people can experience these questions together, with a real agent in the loop.

For a limited time, a room, a system and a set of humans share a task: to see what happens when a non-human authority issues instructions and each person decides how to respond. The traces of that encounter become fossils and memories.

The philosophical claim of HCP is modest and precise. In order to understand our emerging relationship with agentic systems, it is not enough to theorize about capability or control from a distance. It is necessary to inhabit the choreography between instruction and interpretation, to feel the pause before action, and to notice what the body chooses to do when the room moves.