



LAiRP 2026

Research on vs with AI (class 2)

DSI, Zurich
19.10.2026

Phones out · code 79355

roompulse.borant.eu · code 79355

- *Donut Run: the only part of the afternoon where you can stop thinking*



scan to join

Welcome to the LAiRP

Dragon's Lair (Cinematronics, 1983)

- *The most beautiful game anyone had seen (back then at least). Animated by Don Bluth, ex-Disney, playing off a laserdisc.*
- *You weren't really controlling the knight. The animation was already drawn: you pressed the right direction at a handful of exact moments.*
- *Miss one, and you watched a beautifully animated death.*

A lot of research with AI works like that. The output is gorgeous, it moves on its own, and it feels like you're steering. You're choosing at a few moments. **This course is about finding those moments, and being awake when they come.**

Dragon's Lair, Cinematronics 1983 (animation: Don Bluth Productions)

Two questions · code 79355

roompulse.borant.eu · code 79355

- Which of these would you call research? (tick all that apply)
- For your own PhD: how clearly could you say whether AI is your object, your instrument, or both? (1–5)



scan to join



Object or instrument | How it breaks | Both, and your own



Giovanni Spitalè, PhD
LAI RP, Zurich, 19.10.2026



Before any definition

When you open an LLM... are you studying the model, or are you using the model to study something else?

Take 20 seconds

- *What changes if you misidentify which one you are doing?*
- *Where is the boundary in your own work?*

Spitale 2025, Doing research with AI (lecture, IBME Zurich) - giovannispitale.net

Three ways AI can be in your research

Research ON AI

AI is the object.

- *How it behaves, persuades, classifies, discriminates, simulates, fails*
- *Model behaviour, bias, safety, evaluation, failure modes*

Research WITH AI

AI is the instrument.

- *It sits inside your pipeline*
- *Literature mapping, transcription, coding, classification, simulation, drafting*

BOTH

AI as an instrument, in order to study AI.

- *More common than it sounds*
- *And the interesting case: this afternoon is mostly about it*

Same model, same API call: two different sciences

	Research ON AI	Research WITH AI
What threatens validity	prompt sensitivity, model version and drift, sampling temperature, contamination, whether your prompts resemble real use	fabricated sources, coding drift, anchoring, non-reproducibility, missing provenance
What the ethics are about	what you make the model produce (dual use), who the outputs could harm	whose data you send where, who is accountable for what the model concluded
What you will have to report	the model, and how you used it	the model, how you used it, and where it ran and what it saw

What goes into that report, and why, is Michele's session on Tuesday afternoon.

Exercise A · Eight studies, two axes

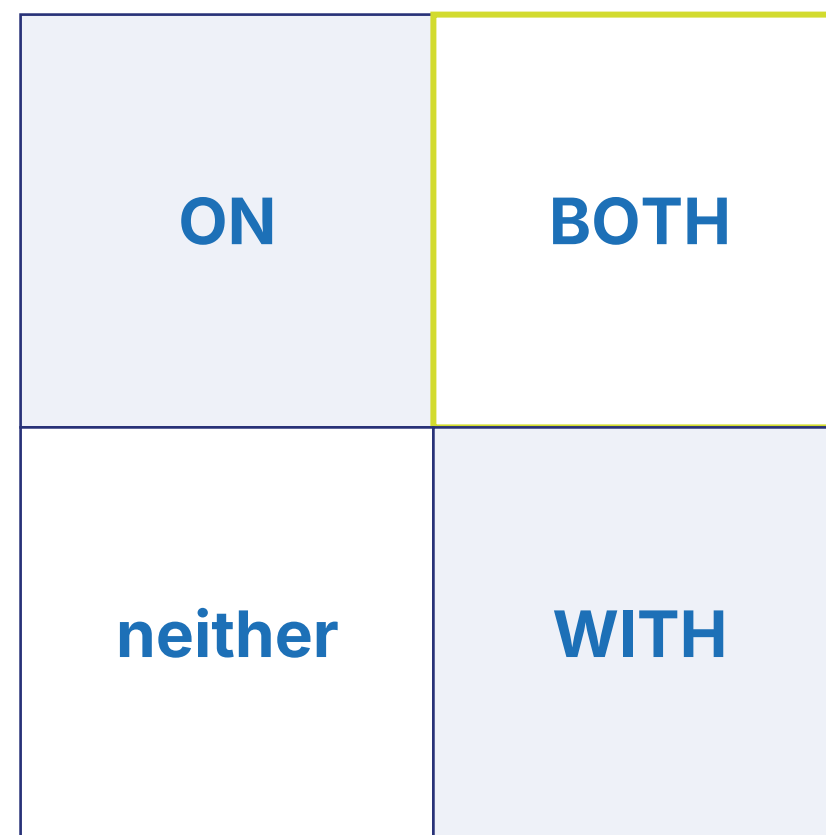
Round 1 · on your own

- Open the study sheet (QR). 3 minutes of silent reading.
- Then, for each study, place one dot on your phone.
- Read to the end of each description.

↑ AI is the object

Round 2 · in groups

- The two studies where the room's dots spread most.
- Agree where each goes, or write both positions.
- Name one threat to validity that follows from it.



Materials

→ AI is the instrument

A1, A5, A8 and A4 are real studies: full references at the end

Place each study · code 79355

roompulse.borant.eu · code 79355

- A1 · GPT-3 vs humans on Twitter
- A2 · Oral history of a factory closure
- A3 · Same CV, different name
- A4 · 249 patient interviews and a dictionary
- A5 · How Google invents things
- A6 · Virtual patients
- A7 · A pilot study without participants
- A8 · Who wrote it?



scan to join

Your project group, until Wednesday

Turing

Lovelace

Hopper

Shannon

RoomPulse puts you in a group: tap once.

These are your groups for everything that follows: **Exercise B, Exercise C, all of tomorrow morning, and the presentation on Wednesday.**

Mixed on purpose: IMHO a workflow is better designed by people who see different kinds of data and constraints.

Exercise A, round 2 · The two most contested

As a group, 10 minutes

- *Take the two studies where the dots spread most.*
- *Agree where each goes. If you can't, write down both positions.*
- *Name one threat to validity that follows from where you put it.*

Then

- *One person reports: 60 seconds per group.*

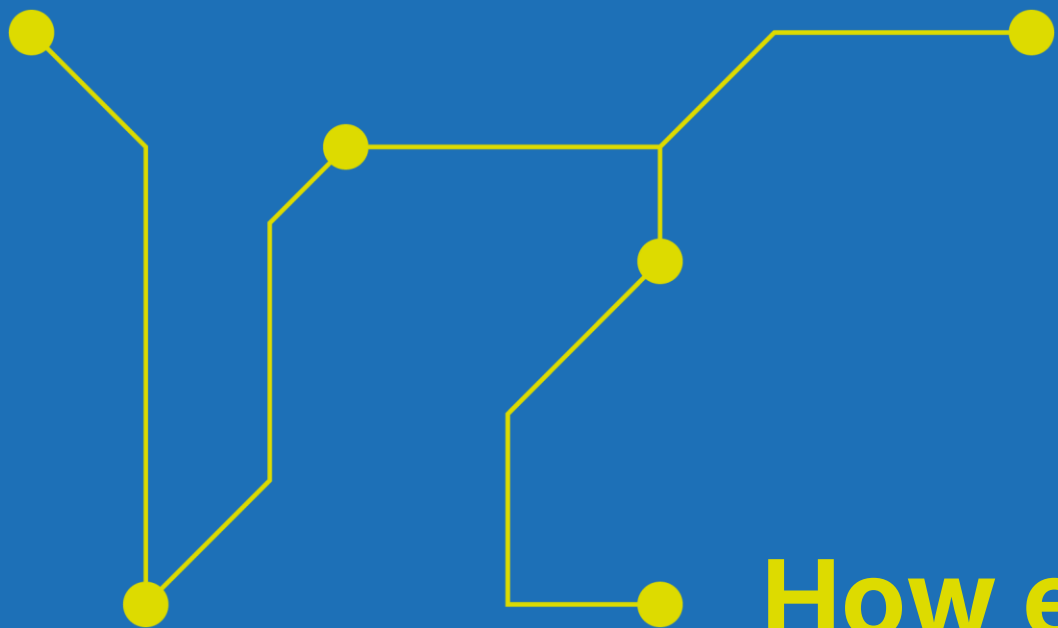
Disagreement is fine. Write down both positions, and why each of you holds yours.

What just happened

The disagreement wasn't about the facts of the study: you all read the same text. It was about what the AI is for in the design. That's the thing you have to be able to say about your own work, and most papers don't say it.

Break · 10 minutes

Sit with your group when you come back.



How each one breaks

Object or instrument | How it breaks | Both, and your own

Research on AI is a method: one lens, three questions

1**Unit of analysis***what is one observation?***2****What is manipulated***and what is held fixed?***3****What could make it false***the threat to validity*

Every study on the next two slides went through these five steps. Look for where a human checked, and where nobody did.

Five studies through the lens (1/2)

Study	Unit	Manipulated	What could make it false
GPT-3 (dis)informs us better than humans (Spitale, Biller-Andorno, Germani 2023, Sci Adv)	a tweet, judged by a person (697 participants)	author (human / GPT-3) × truth (true / false)	GPT-3 is no longer the model anyone uses
Emotional prompting (Vinay et al. 2025, Front AI)	one of 19,800 generated posts	politeness of the request × model × system role	does anyone prompt like that? and a 'helpful assistant' system role makes the effect disappear
Source framing (Germani & Spitale 2025, Sci Adv)	an agreement score (192,000 of them)	the declared source of the same statement	the models also wrote the statements: bias in judging, or in writing?

Spitale, Biller-Andorno & Germani 2023 · Vinay et al. 2025 · Germani & Spitale 2025 (full references at the end)

Five studies through the lens (2/2)

Study	Unit	Manipulated	What could make it false
Assisted-suicide case reports (Spitale, Schneider et al. 2023, Front AI)	one of 72 Dutch case reports	three classifiers on the committee's verdict: logistic regression, zero-shot NLI, GPT-4	class imbalance: 56 yes, 16 no, so always saying 'yes' scores 78%; and the prompt: 'does everyone agree' turned every contested case into a no
How Google AI Overview confabulates (Germani & Spitale, in progress)	one of 140 overviews	kind of invented entity, 10 repetitions each	the coding itself is done by an LLM

Ethics, one beat: research on AI tends to end with a detector. A tool that hands down a verdict replaces the judgement it was meant to educate: the paternalism trap.

Spitale et al. 2023 · Spitale & Germani 2026, preprint - <https://doi.org/10.5281/zenodo.20321941>

Exercise B · Kill the study

Your card describes one study. Be Reviewer 2.

- List every way the result could be wrong, misleading, or impossible to repeat.
- Be specific: not 'bias', but **which** bias, entering **where**.
- Pick the one that worries you most: what would you do about it?
- 20 minutes. Last 3: one person submits on RoomPulse, one entry per threat.



Cards

Card ON

Turing · Hopper

Emotional prompting: 19,800 posts written 'as Sam'

Card WITH

Lovelace · Shannon

An LLM codes sixty interviews

Card ON: Vinay, Spitale, Biller-Andorno & Germani 2025 - <https://doi.org/10.3389/frai.2025.1543603>

Submit your threats · code 79355

roompulse.borant.eu · code 79355

- *Threats: the ON study (emotional prompting)*
- *Threats: the WITH study (an LLM codes sixty interviews)*



scan to join

Two epistemologies

ON: emotional prompting

- the models used are retired: not replicable as run
- prompt and system role decide the result
- the manipulation itself was written by gpt-3.5
- ecological validity: does anyone prompt like this?
- dual use: the paper is also a recipe

WITH: an LLM codes interviews

- quotes paraphrased, or invented, presented as verbatim
- codebook drift: the codebook applied $\neq$ the one declared
- synonym explosion: a new code for almost every excerpt
- agreement with humans measured on a convenient subset, or not at all
- coding in batches changes the result

Left: did we measure the model correctly? Right: did the model measure our data correctly? **A lens you can't inspect is a lens you can't trust.**

Three numbers for the WITH side

473

codes for 555 coded segments

First exploratory run of the AI Overview study, empty codebook, vague study description. Nothing was wrong with the model; everything was wrong with what it was given.

β -0.41

batched vs sequential coding

Coding segments in batches significantly lowered agreement with human consensus ($p = 0.007$). Same models, same task, different structure.

0.75 vs 0.42

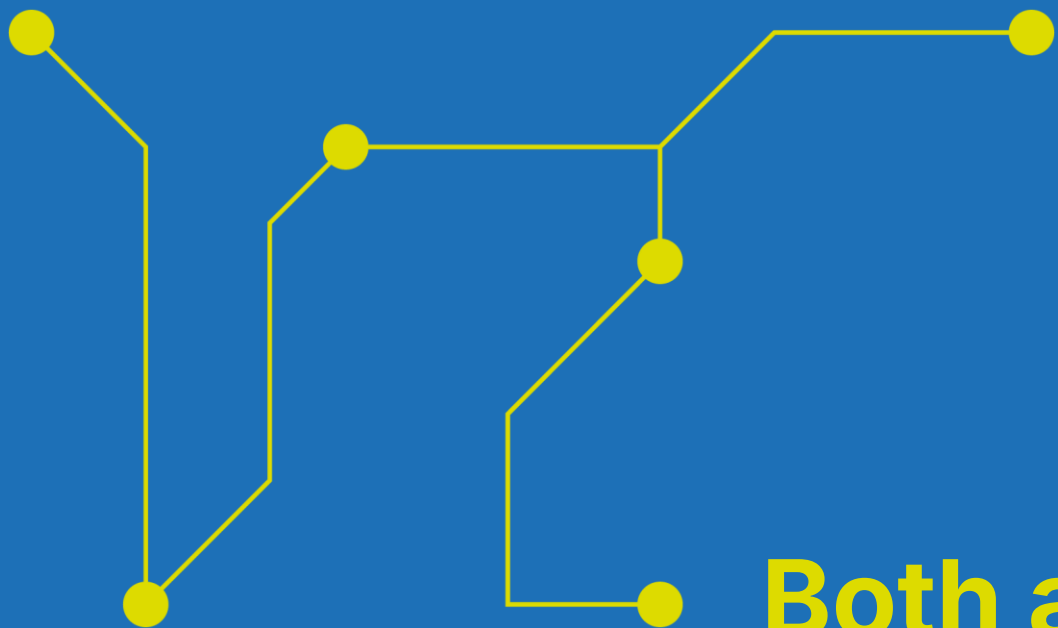
κ : LLM-LLM vs human-human

The models agreed with each other more than humans did, and their disagreements were more systematically patterned. Consistent is not the same as right.

Rush et al. 2026, Medical Teacher - <https://doi.org/10.1080/0142159x.2026.2721362>

Break · 10 minutes

Questions for me: RoomPulse, any time from after the break.



Both at once, and your own

Object or instrument | How it breaks | Both, and your own

Four things you can ask an LLM to do

Compressors

summaries · literature triage · first-pass themes

Extractors

transcription · speaker turns · raw text into structured data

Expanders

brainstorming · counterarguments · alternative designs

Formalisers

argument structures · templates · natural language into R or Python

Tomorrow morning is built on these four.

Spitale 2025, Doing research with AI (lecture, IBME Zurich) - giovannispitale.net

'Can AI do qualitative coding?' is a badly posed question

Like asking whether statistics can be used for research: the answer depends on what surrounds it.

A live example

A 2025 manifesto, five authors and 419 signatories, rejects generative AI for reflexive qualitative research.

Its evidence base is overwhelmingly ad-hoc chatbot use. Its conclusion is about AI as such.

They banned a different object from the one they observed.

Jowsey, Braun, Clarke, Lupton & Fine 2025, *Qualitative Inquiry* - <https://doi.org/10.1177/10778004251401851>

The apparatus, not the model

Property of the tool	What it prevents
Prompts inspectable and versioned	an analysis that cannot be repeated; disclosure as anecdote
Runs versioned and re-runnable	'the AI said' as an event instead of a method
Excerpts anchored to literal spans of the source	fabricated quotes: not unlikely, structurally impossible
The codebook as a declared constraint	silent drift between the codebook you report and the one applied
Explicit human gates on accepting and rejecting	a tool that hands down verdicts instead of proposals

The last row is the moment in the lair where you have to press the button. Everything above it is there so you know when the moment has come.

What each row looks like when someone builds it

LIVE DEMO · AutoCode, 3 minutes

I'm not showing you a product: I'm showing you what each row of the table looks like when someone builds it.

Human or machine?

“And so, I mean, we found - we found the lifestyle that did work. I mean, my partner's amazing. He was constantly washing bottles, or he could do a lot of the feedings, which was really good. I was just like, it was so many emotions at the beginning with having to deal with a baby with a cleft palate because it was - I mean - you just, you have to figure it out for yourself.”

Published patient interview · prenatal testing · one turn

Code 1

Relationship with partner

Codebook: demographic and descriptive details of relationship

Code 2

Work and Pregnancy

Codebook: references to working during pregnancy, actual or contemplated

Code 3

Health variations – Impact on family life

Codebook: the influence of health variations on the larger family

Code 4

Need for/missing information

Codebook: information that is not available but should be

Interviews: Fang et al. 2023, BMC Med Ethics; Fang et al. 2025, Narrat Inq Bioeth

Who wrote it? · code 79355

roompulse.borant.eu · code 79355

- Code 1: human researcher or AI?
- Code 2: human researcher or AI?
- Code 3: human researcher or AI?
- Code 4: human researcher or AI?



scan to join

What were you using to decide?

1 · Relationship with partner

AI

2 · Work and Pregnancy

Human

3 · Health variations – Impact on family life

AI

4 · Need for/missing information

Human

There was no style to go on: same codebook, same labels. All you could judge was whether a code fits the text. That is the right question, and it doesn't tell you who coded.

Why the human's codes look off here: the human coded the passage together with the next turn ("I knew more than the doctors for like what's practical"). Same codes on both. A different unit, not a careless coder.

Four items prove nothing, and this isn't the study. Whether a code is right about the text is the question the manifesto skips, and the one a real test has to ask.

The reflexive case: using the tool to study the tool

1

Code the same material by hand, and with AutoCode

2

Mix the two

3

Ask experienced qualitative researchers, ideally signatories of the manifesto, which is which

With and on at the same time. And the instrument I'm studying is the instrument I'm using. If you can tell me where the slippage is, you understood the afternoon.

Exercise C · Your research, two ways

In your group, in the project sheet (QR)

- *Pick one topic close to at least one person's PhD.*
- *Write a question where AI is the object, and one where it is the instrument.*
- *Choose one to take to the project, and say why.*
- *What could sink it? (Look back at Exercise B.)*
- *15 minutes on the sheet, then 5 minutes on RoomPulse: three steps, one chain.*

What's the question you didn't pick, and why not? It's often the better project.



Project
sheet

One chain per group · code 79355

roompulse.borant.eu · code 79355

- *Step 1 · The question we're taking to the project*
- *Step 2 · Why AI is the object, the instrument, or both here*
- *Step 3 · The threat that could sink it*



scan to join

Before you go · code 79355

roompulse.borant.eu · code 79355

- *Same question as before: how clearly could you say whether AI is your object, your instrument, or both?*
- *Questions: the top two now, the rest tomorrow at 9:30*



scan to join

Tonight · about 20 minutes, on Teams

Your question, step by step

On your own, take your group's question and write how you would study it, in five or six steps. For each step:

- *what you would ask a tool to do (or nothing)*
- *which tool you would use **tonight**: the one you'd really reach for, not the right one*
- *the level of the data that goes into it*

Then one line: the step where this is most likely to go wrong.

Due Tuesday 08:00

No account to create, nothing to install.

Can't do it tonight? Come anyway: you'll do it aloud with your group.

Tomorrow at 9:30

Your question becomes a sequence of tools.

Bring nothing but what you write tonight.



Choose 'Class 2' on the first page

One minute, anonymous: how was this afternoon?

References

- Fang MT, Germani F, Spitale G, Wüscher S, Kunz L, Biller-Andorno N (2023). Women's experiences with non-invasive prenatal testing in Switzerland: a qualitative analysis. *BMC Medical Ethics*, 24(1), 85. doi:10.1186/s12910-023-00964-3
- Fang MT, Kunz L, Spitale G, Wüscher S, Germani F, Biller-Andorno N (2025). Women's perceptions of the medicalization of pregnancy and their preferred models of care: a qualitative analysis. *Narrative Inquiry in Bioethics*, 15(1), 53–68. doi:10.1353/nib.2025.a957205
- Germani F, Spitale G (2025). Source framing triggers systematic bias in large language models. *Science Advances*, 11(45), eadz2924. doi:10.1126/sciadv.adz2924
- Jowsey T, Braun V, Clarke V, Lupton D, Fine M (2025). We reject the use of generative artificial intelligence for reflexive qualitative research. *Qualitative Inquiry*. doi:10.1177/10778004251401851
- Rush E, Karim MN, Yacu GS, et al. (2026). Evaluating hybrid human-LLM coding workflows for qualitative research in medical education: a generalizability study. *Medical Teacher*. doi:10.1080/0142159x.2026.2721362
- Spitale G (2025). *Doing research with AI* [lecture slides]. IBME, University of Zurich, 10 December 2025. giovannispitale.net
- Spitale G, Biller-Andorno N, Germani F (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26), eadh1850. doi:10.1126/sciadv.adh1850
- Spitale G, Germani F (2026). Misinformation, disinformation, and autonomy: against the epistemic paternalism trap [preprint]. *Zenodo*. doi:10.5281/zenodo.20321941
- Spitale G, Schneider G, Germani F, Biller-Andorno N (2023). Exploring the role of AI in classifying, analyzing, and generating case reports on assisted suicide cases: feasibility and ethical implications. *Frontiers in Artificial Intelligence*, 6, 1328865. doi:10.3389/frai.2023.1328865
- Vinay R, Spitale G, Biller-Andorno N, Germani F (2025). Emotional prompting amplifies disinformation generation in AI large language models. *Frontiers in Artificial Intelligence*, 8, 1543603. doi:10.3389/frai.2025.1543603

Work in progress, shown in class, not yet citable: the AI Overview confabulation study (Germani & Spitale); the patient-derived taxonomy of healthcare harm (Spitale, Naishtat, Ritzmann, Biller-Andorno); the AI Coding Turing Test (Spitale et al.).



Let's keep talking

giovanni.spitale@ibme.uzh.ch



Slides and materials



University of
Zurich ^{UZH}

Institute of Biomedical Ethics
and History of Medicine

