AI Governance

Which system, under which conditions, with which safeguards, for which data

DSI Excellence Bootcamp, Zurich

03.09.2026

Join at borant.eu/roompulse · code 81197

ITE LAB
Information Technology & Experimental Ethics

University of Zurich[UZH]
Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

Educated people introduce themselves

Giovanni Spitale, PhD
Institute of Biomedical Ethics and History of Medicine (IBME)University of Zurich, Switzerland.
Co-director of the **ITE Lab** (Information, Technology & Experimental Ethics Lab) bridging empirical research, public engagement, and creative methods.

Other fancy stuff:
- WHO rapporteur (ethical guidance for infodemic management)
- TEDx speaker
- Open Science Ambassador @ UZH
- Bioetica da Bar
- Paragliding pilot, nerd, big fan of cows

www.giovannispitale.net - www.borant.eu/tools/

Introduction

Introduction | Data levels | Case classification | AI tools | Capability gap

Information Technology & Experimental Ethics

University of Zurich[UZH]

Institute of Biomedical Ethics
and History of Medicine

Giovanni Spitale, PhD
DSI Excellence Bootcamp, Zurich, 03.09.2026

~~Let me scare you AF~~

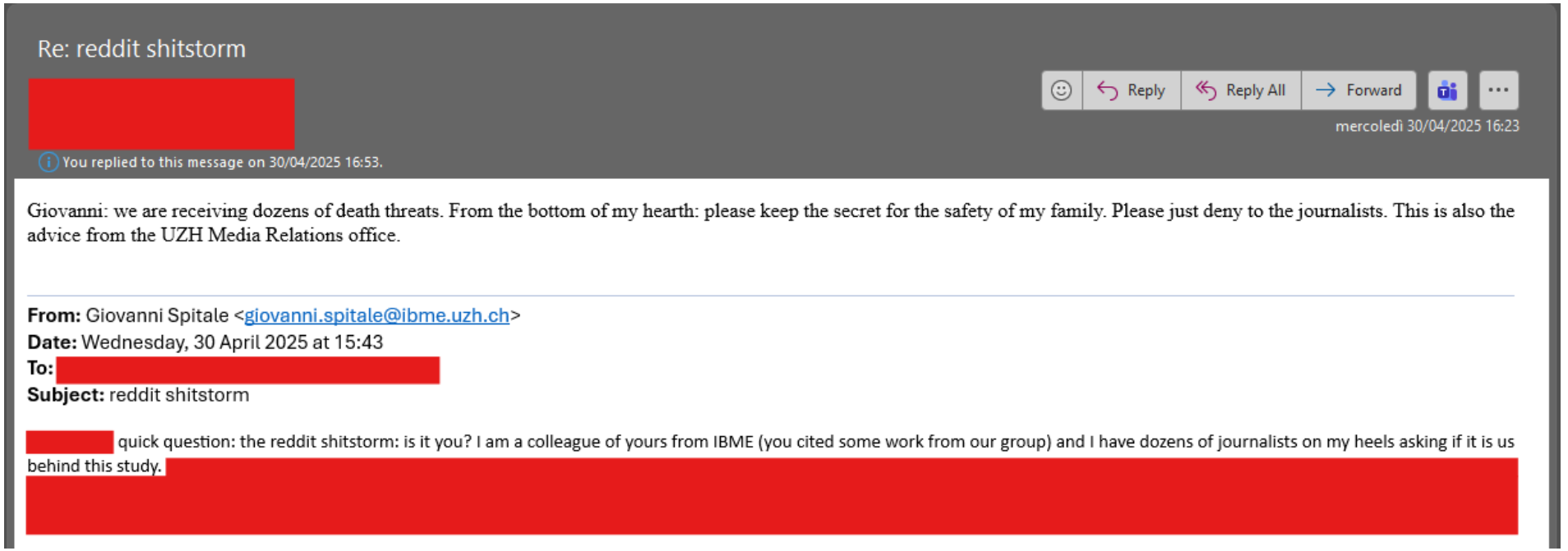
Why it's important

→ Researchers infiltrated r/ChangeMyView with AI-generated profiles posing as real users.
→ Bots engaged in live debates for months to measure the persuasive power of LLMs "in the wild."
→ Participants were unaware; no informed consent or debriefing plan was in place.

Responsible research with SoMe and health data - https://www.giovannispitale.net/responsible-research-with-social-media-and-health-data/

~~Let me scare you AF~~

Why it's important



Responsible research with SoMe and health data - https://www.giovannispitale.net/responsible-research-with-social-media-and-health-data/

→ RoomPulse

Code 81197

ITE LAB Information Technology & Experimental Ethics

University of Zurich[UZH]

Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

~~AI: yes or no~~

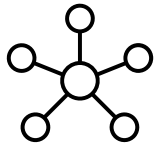
"AI: yes or no" is unanswerable, which is why nobody obeys it.

LAB
Information Technology & Experimental Ethics

University of Zurich[UZH]
Institute of Biomedical Ethics and History of Medicine

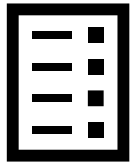
Giovanni Spitale, PhD

The real question has four variables, and all four are decidable

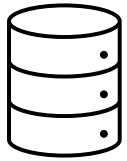


Which system

With which safeguards



Under which conditions



For which type of data

Today's plan

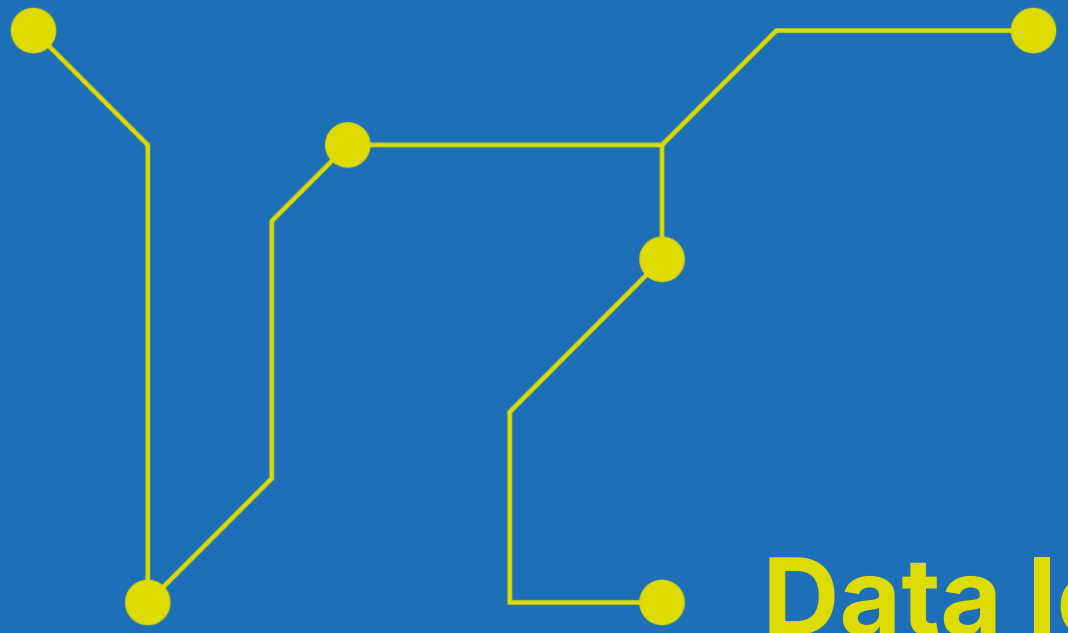
The **data** has a **level**, the **tool** has a **ceiling, the decision is a match.**

- Level: not a feeling about how sensitive something seems. A defined level, in a directive, with law behind it.

- Ceiling: the highest level a tool may touch. Contractual and legal, which is why two tools that behave identically can have different ceilings.

- Match: ceiling at or above green light level.

If you remember nothing else from today, remember these three lines.



Data levels

Introduction | Data levels | Case classification | AI tools | Capability gap

Information Technology & Experimental Ethics

University of Zurich[UZH]

Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

DSI Excellence Bootcamp, Zurich, 03.09.2026

Where the levels come from

Weisung zur Klassifizierung von Informationen (UZH), in force since 01.09.2023, based on cantonal law: IDG, IVSV

Objective	Protects	Failure looks like
Confidentiality	The people the data is about	A participant re-identified from something told in confidence during one interview
Integrity	Everyone downstream who relies on the findings	A corrupted or altered dataset becomes a paper and informs a policy
Availability	Research and researchers' time	Data lost or locked away, so their contribution to science amounts to 0
Accountability	The possibility of correction	Nobody can reconstruct who did what so nothing can be audited

Weisung zur Klassifizierung von Informationen (UZH, 21.05.2021, in force 01.09.2023) - https://www.rud.uzh.ch/dam/jcr:a86f65ae-d114-4e57-bff7-47b024eb72f3/Weisung%20zur%20Klassifizierung%20von%20Informationen.pdf

→ RoomPulse (six quick ones)

Code 81197

Öffentlich (open); intern (internal); vertraulich (confidential); geheim (secret)

LAB
Information Technology & Experimental Ethics

University of Zurich[UZH]
Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

Three rules

Three rules cover almost everything you just got wrong:

1. An **unmarked document** is at least **intern.** There is **no "unclassified".**
2. Until publication, everything is **at least intern.** Including your own work. *Especially your own work.*
3. **Classification is inherited** across whole information holdings. One sensitive item pulls its container up with it.

Weisung zur Klassifizierung von Informationen (S6)

Four data levels

Level		What it covers
Öffentlich	Public	Published information, or information cleared for publication. E.g.: a published paper or preprint
Intern	Internal	Unpublished work, internal documents. The default for anything not explicitly marked. E.g.: an email from your institute's mailing list
Vertraulich	Confidential	Personnel files, student data, personal research data, lab notebooks, logs…
Geheim	Secret	Professional secrecy (art. 321/321bis Swiss criminal code: patient data, human research in medicine and psychology; contractual NDAs; judicial files; …)

Geheim: **careful with this one.** It isn't "more confidential than vertraulich" - it's a different regime, and you get into it in three ways: **professional secrecy** under article 321, **human research** in medicine and psychology, or **a contract**. Notice that none of those three is about how sensitive the data feels.

StGB/CP, SR 311.0, art. 321 and 321bis - fedlex.admin.ch/eli/cc/54/757_781_799/de
LRUm/HFG/HRA, SR 810.30 - https://www.fedlex.admin.ch/eli/cc/2013/617/de

Geheim / secret

The same interview, word for word, can be vertraulich in one project and geheim in another. What changes isn't the text. It's **who collected it, and under what obligation.**

Research data with identifiable participants is vertraulich, **unless it comes out of a clinical study, and then it moves.**

So the question you'll be asking on your card is not “how sensitive is this?”. **It’s “under what obligation did I get it?”.**

So I can't do research on clinical data at all?"

- Yes you can, clinical research is a thing that happens, lawfully, every day in this university.
- The route runs through the Human Research Act and an ethics committee, and art. 321bis para. 2 is the hinge that connects criminal-law secrecy to it. What art. 34 covers specifically is the exceptional case of further use without consent.

StGB/CP, SR 311.0, art. 321 and 321bis - fedlex.admin.ch/eli/cc/54/757_781_799/de
LRUm/HFG/HRA, SR 810.30 - https://www.fedlex.admin.ch/eli/cc/2013/617/de

Break (15')

LAB
Information Technology & Experimental Ethics

University of Zurich[UZH]
Institute of Biomedical Ethics
and History of Medicine

Giovanni Spitale, PhD

Case classification

Introduction | Data levels | Case classification | AI tools | Capability gap

Information Technology & Experimental Ethics

University of Zurich[UZH]

Institute of Biomedical Ethics
and History of Medicine

Giovanni Spitale, PhD

DSI Excellence Bootcamp, Zurich, 03.09.2026

→ RoomPulse

Code 81197

LAB
Information Technology & Experimental Ethics

University of Zurich[UZH]
Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

Exercise 1: classify the case

Six artefacts, four levels, one line of justification each.

Your card describes a research project and six data artefacts.

For each artefact:

1. assign one of the four levels **(open, internal, confidential, secret)**
2. write **one line** saying why
3. You have **14 minutes.**
4. Disagreement inside the group is fine, note it, you will be reporting afterwards

Download the handouts / material ->

ITE LAB Information Technology & Experimental Ethics

University of Zurich[UZH]

Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

Case A: digital health

Group report

You are running a twelve-month **cohort study on adherence to a smartphone monitoring app** among patients with a chronic condition. Mixed methods: **app telemetry plus focus groups.**
Your artefacts:

1. Open-access literature on adherence
2. Your study protocol, written, not yet published
3. Pseudonymised telemetry export from the app (step counts, timestamps, device IDs)
4. Scanned signed consent forms, with names and signatures
5. Focus-group transcripts in which patients describe symptoms and treatment
6. Your draft manuscript

Explain the mapping artefact -> level

60 seconds

Case A: digital health

The reading

Artefact	Level
Literature	Public
Study protocol	Internal
Pseudonymized telemetry	Confidential ⚠️
Signed consent forms	Confidential
Focus group transcripts	Secret ⚠️
Draft manuscript	Internal

Case B: case law and digital justice

Group report

You are comparing how **criminal courts in three jurisdictions** handle evidence produced by algorithmic systems (e.g. risk-assessment scores, automated forensic matching). Mixed methods: **published judgments, unpublished case files obtained under a research agreement with one court, and interviews with prosecutors.**
Your artefacts:

1. Published judgments from public databases
2. The statutory corpus
3. Court files that are not publicly available, obtained under a research agreement
4. Your interview notes from conversations with prosecutors
5. A manuscript by another researcher that you are currently peer-reviewing
6. Your own draft article

Explain the mapping artefact -> level

60 seconds

Case B: case law and digital justice

The reading

Artefact	Level
Published judgments	Public
Statutory corpus	Public
Non-public court files	Secret
Interview notes, prosecutors	Confidential ⚠️
Manuscript under review	Confidential (obligation you accepted)
Draft manuscript	Internal

Case C: accessibility and interaction design

Group report

You are testing a **screen-reader prototype** you built, with blind and low-vision participants, benchmarked against a public dataset. Sessions are **screen- and video-recorded** and followed by a **short questionnaire;** you have presented early results and are drafting a follow-up grant.
Your artefacts:

1. A public benchmark dataset
2. The prototype source code - yours, not yet released
3. Screen and video recordings of the test sessions, including participants' faces and voices
4. Post-session questionnaires with demographic data
5. Your draft grant proposal for the follow-up project
6. Conference slides you have already presented

Explain the mapping artefact -> level

60 seconds

Case C: accessibility and interaction design

The reading

Artefact	Level
Benchmark dataset	Public
Prototype code	Internal
Session recordings	Secret / confidential ⚠️
Demographic questionnaires	Confidential
Grant proposal	Internal
Conference slides	Public

Case D: Disinformation and content moderation

Group report

You are studying **how a large platform justifies its moderation decisions.** Mixed methods: **public posts collected through the platform API and coded by a small team against your codebook, plus interviews with moderators granted under NDA.** The project is externally funded.
Your artefacts:

1. Public posts collected through the platform API, under its terms of service
2. Your codebook
3. Interviews with content moderators, conducted under NDA
4. The annotated dataset (posts, codes, coder identifiers)
5. Your preprint
6. The interim report to your funder

Explain the mapping artefact -> level

60 seconds

Giovanni Spitale, PhD

Case D: Disinformation and content moderation

The reading

Artefact	Level
Public posts via API	Public ⚠️
Codebook	Internal
Moderator interviews (NDA)	Secret
Annotated dataset	Confidential
Preprint	Public
Report to funder	Internal ⚠️

Giovanni Spitale, PhD

Case E: Educational research

Group report

You are **evaluating a new teaching format** you introduced in one seminar group of a large undergraduate course. Mixed methods: **the group's coursework and exam scripts, free-text course evaluations, and recordings of the sessions.** You are writing it up for the faculty.
Your artefacts:

1. The published curriculum and syllabus
2. Student essays and exam scripts from the intervention
3. Course evaluation free-text responses - the cohort is small
4. Recordings of the lectures, including student questions
5. Your own teaching materials
6. Your draft report to the faculty

Explain the mapping artefact -> level

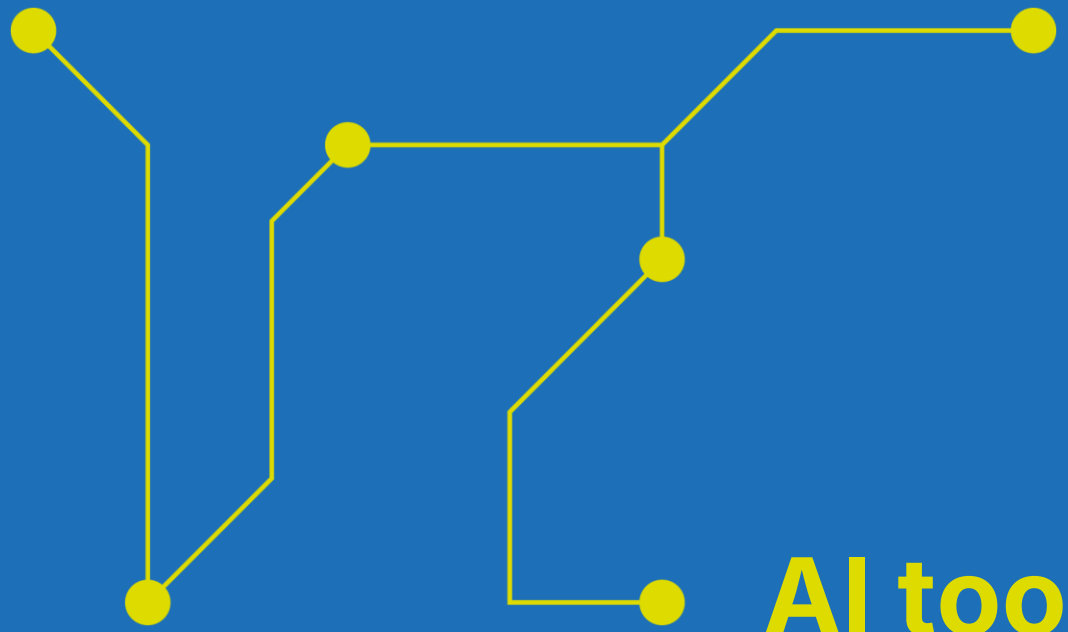
60 seconds

Giovanni Spitale, PhD

Case E: Educational research

The reading

Artefact	Level
Published CV	Public
Student essays and exams	Confidential
Free text evaluations	Confidential
Lecture recordings	Confidential
Own teaching materials	Internal ⚠️
Draft report to faculty	Internal



AI tools

Introduction | Data levels | Case classification | AI tools | Capability gap

ITE LAB Information Technology & Experimental Ethics

University of Zurich[UZH]

Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

DSI Excellence Bootcamp, Zurich, 03.09.2026

Categories of use

- Compressors
- Extractors
- Expanders
- Formalizers

Every tool has a ceiling, and the free ones are at the bottom

Tool	Level (up to)	Hosting
Free AI tools (ChatGPT, Claude, Deepl, ...)	Public	Worldwide cloud
M365 Copilot **chat** (**basic**, UZH account)	Internal	EU cloud
M365 Copilot **premium** (UZH account)	Confidential (web search disabled)	EU cloud
Deepl Pro Advanced	Confidential	EU cloud
GitHub Copilot	Internal	Worldwide cloud
Science**Cloud** (GPU research)	Confidential (depends on usage context)	On premises
Science**Cluster** (GPU research)	Internal	On premises
Supercomputer (GPU research)	Internal (non personal data only)	CH cloud
LM Studio + Mistral	Confidential	On device

Things change fast!

e.g. Claude, Claude Code currently in evaluation; LLM Proxy planned (12/2026); ChatGPT planned (09/2026); ... - https://www.zi.uzh.ch/de/staff/software-elearning/ai-tools-und-services.html

Check these sources:

- Übersicht KI-Tools und Services, ZI UZH - https://www.zi.uzh.ch/de/staff/software-elearning/ai-tools-und-services.html
- Microsoft Copilot richtig und sicher nutzen (ZI blog, 12.05.2026) - https://www.uzh.ch/blog/zi/2026/05/12/microsoft-copilot-richtig-und-sicher-nutzen/
- Merkblatt Copilot (ZI, PDF) - https://www.zi.uzh.ch/dam/jcr:402ddecd-9015-4fb1-b8e5-5e5077d9ad06/Merkblatt_ZI_Copilot.pdf

The line most of you already crossed:

Free AI tools (ChatGPT, Claude, DeepL, ...) are for public data only.

For secret data there is no AI tool at all.

LAB
Information Technology & Experimental Ethics

University of Zurich[UZH]
Institute of Biomedical Ethics and History of Medicine

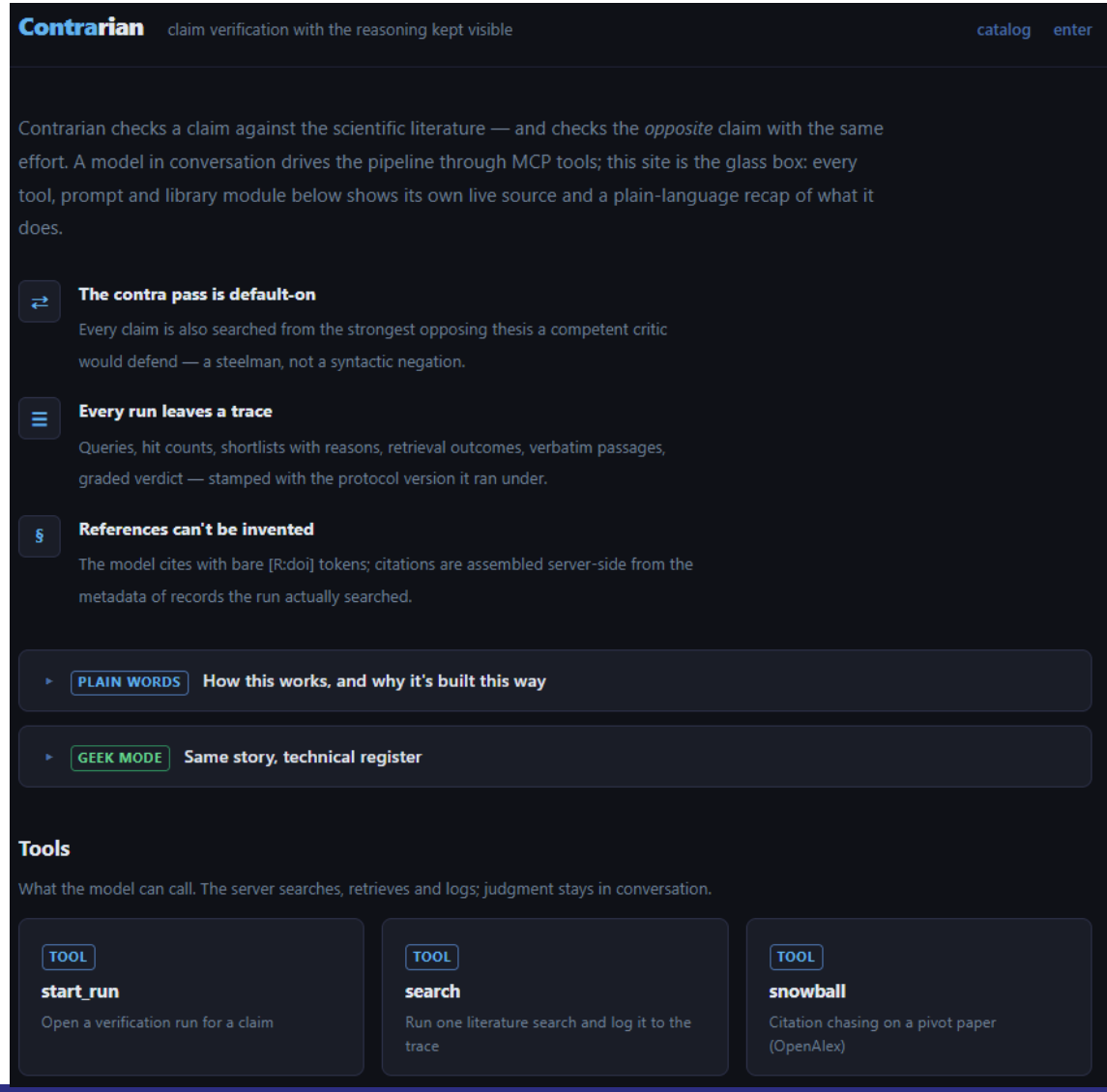
Giovanni Spitale, PhD

Case study: Contrarian

The question is never "what level is this tool“; it is "what level is this flow, at this point".

Contrarian: a tool that takes a claim, searches PubMed / Europe PMC / OpenAlex for evidence **for and against** it, retrieves full texts, and returns a graded verdict.

Q: What level is it working at?
Q: Could you run it against a commercial cloud model?



Contrarian – www.contrarian.borant.eu

Exercise 2: Pick a tool for each artefact

(and find the one where you cannot).

Same cases (A-E). Same groups.

Task 1: for each artefact: which tool, for what, **with what safeguard**.

Task 2: you might find at least one artefact where the answer is **no AI tool at all**.
What do you actually do?

15 minutes + presentation (60” per group)

Download the handouts / material ->

ITE LAB Information Technology & Experimental Ethics

University of Zurich[UZH]

Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

Case A: digital health

Artefact	Level
Literature	Public
Study protocol	Internal
Pseudonymized telemetry	Confidential ⚠️
Signed consent forms	Confidential
Focus group transcripts	Secret ⚠️
Draft manuscript	Internal

Tool	Level (up to)
Free AI tools (ChatGPT, Claude, Deepl, ...)	Public
M365 Copilot **chat** (**basic**, UZH account)	Internal
M365 Copilot **premium** (UZH account)	Confidential (web search disabled)
Deepl Pro Advanced	Confidential
GitHub Copilot	Internal
Science**Cloud** (GPU research)	Confidential (depends on usage context)
Science**Cluster** (GPU research)	Internal
Supercomputer (GPU research)	Internal (non personal data only)
LM Studio + Mistral	Confidential

Case B: case law and digital justice

Artefact	Level
Published judgments	Public
Statutory corpus	Public
Non-public court files	Secret
Interview notes, prosecutors	Confidential ⚠️
Manuscript under review	Confidential
Draft manuscript	Internal

Tool	Level (up to)
Free AI tools (ChatGPT, Claude, Deepl, ...)	Public
M365 Copilot **chat** (**basic**, UZH account)	Internal
M365 Copilot **premium** (UZH account)	Confidential (web search disabled)
Deepl Pro Advanced	Confidential
GitHub Copilot	Internal
Science**Cloud** (GPU research)	Confidential (depends on usage context)
Science**Cluster** (GPU research)	Internal
Supercomputer (GPU research)	Internal (non personal data only)
LM Studio + Mistral	Confidential

Case C: accessibility and interaction design

Artefact	Level
Benchmark dataset	Public
Prototype code	Internal
Session recordings	Secret / conf. ⚠️
Demographic questionnaires	Confidential
Grant proposal	Internal
Conference slides	Public

Tool	Level (up to)
Free AI tools (ChatGPT, Claude, Deepl, ...)	Public
M365 Copilot **chat** (**basic**, UZH account)	Internal
M365 Copilot **premium** (UZH account)	Confidential (web search disabled)
Deepl Pro Advanced	Confidential
GitHub Copilot	Internal
Science**Cloud** (GPU research)	Confidential (depends on usage context)
Science**Cluster** (GPU research)	Internal
Supercomputer (GPU research)	Internal (non personal data only)
LM Studio + Mistral	Confidential

Case D: Disinformation and content moderation

Artefact	Level
Public posts via API	Public ⚠️
Codebook	Internal
Moderator interviews (NDA)	Secret
Annotated dataset	Confidential
Preprint	Public
Report to funder	Internal ⚠️

Tool	Level (up to)
Free AI tools (ChatGPT, Claude, Deepl, ...)	Public
M365 Copilot **chat** (**basic**, UZH account)	Internal
M365 Copilot **premium** (UZH account)	Confidential (web search disabled)
Deepl Pro Advanced	Confidential
GitHub Copilot	Internal
Science**Cloud** (GPU research)	Confidential (depends on usage context)
Science**Cluster** (GPU research)	Internal
Supercomputer (GPU research)	Internal (non personal data only)
LM Studio + Mistral	Confidential

Case E: Educational research

Artefact	Level
Published CV	Public
Student essays and exams	Confidential
Free text evaluations	Confidential
Lecture recordings	Confidential
Own teaching materials	Internal ⚠️
Draft report to faculty	Internal

Tool	Level (up to)
Free AI tools (ChatGPT, Claude, Deepl, ...)	Public
M365 Copilot **chat** (**basic**, UZH account)	Internal
M365 Copilot **premium** (UZH account)	Confidential (web search disabled)
Deepl Pro Advanced	Confidential
GitHub Copilot	Internal
Science**Cloud** (GPU research)	Confidential (depends on usage context)
Science**Cluster** (GPU research)	Internal
Supercomputer (GPU research)	Internal (non personal data only)
LM Studio + Mistral	Confidential

Break (15')

LAB
Information Technology & Experimental Ethics

University of Zurich[UZH]
Institute of Biomedical Ethics
and History of Medicine

Giovanni Spitale, PhD

The capability gap

Introduction | Data levels | Case classification | AI tools | Capability gap

Information Technology & Experimental Ethics

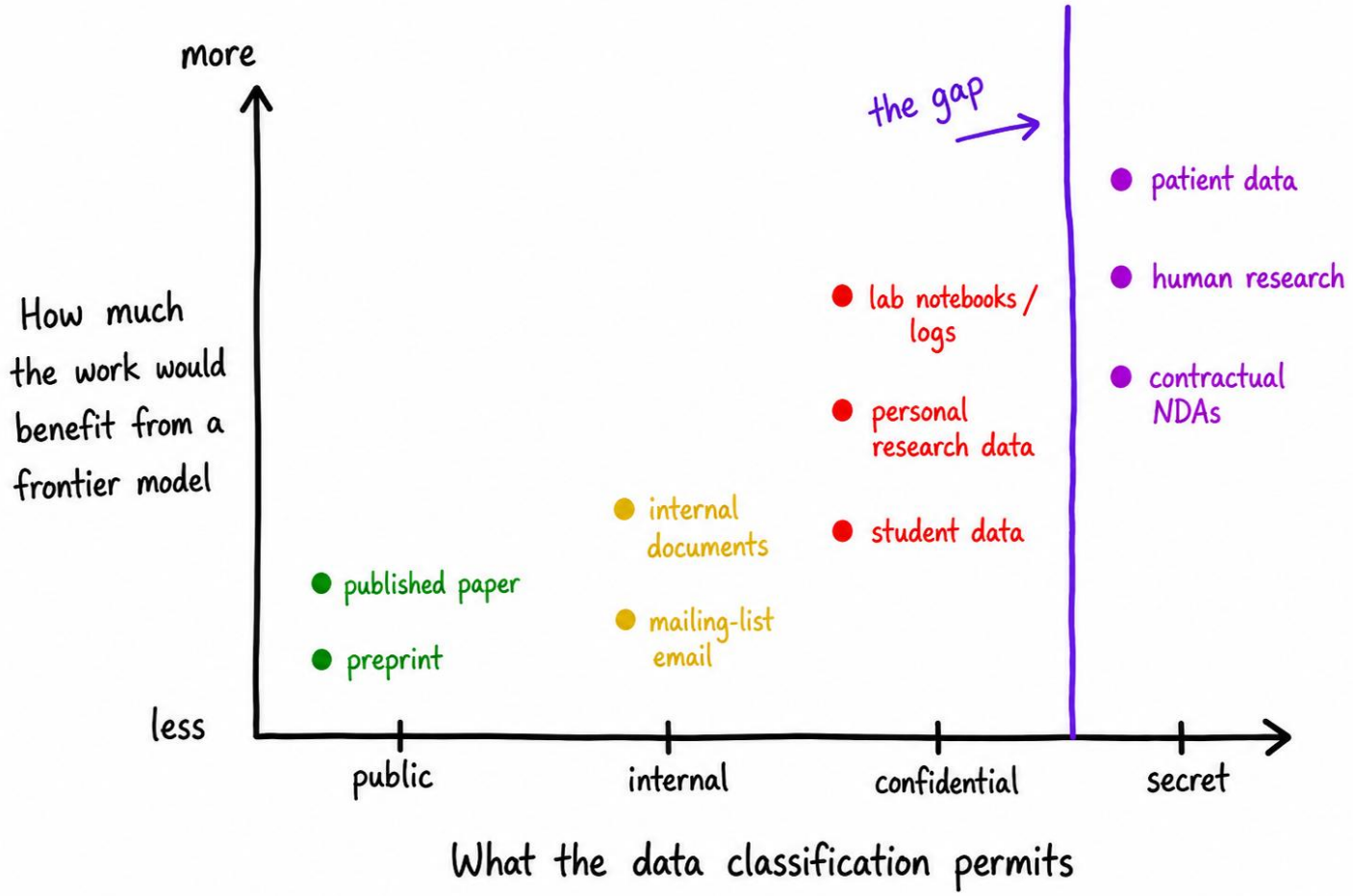
University of Zurich[UZH]

Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

DSI Excellence Bootcamp, Zurich, 03.09.2026

The capability gap...



… explained

The research that most needs frontier models is the research with the weakest tools.

Frontier models live in the cloud. The cloud stops at internal.

So: sensitive data, interview transcripts, evidence synthesis — the work with the most to gain has the least available.

This is the (current) shape of the policy; this is the (current) toolset.

A moving target

The matrix you just learned will be wrong within months. Status (27.08.2026):

Tool	Status	Planned
ChatGPT	Under review	09/26
LLM Proxy (ZI)	Planned	12/26
JupyterHub	Pilot	01/27
Unsloth studio	Planning	03/27
Le Chat / studio (Mistral)	Planning	06/27
Claude / Claude Code	Under review	?
Google AI	Negotiations	?
AWS AI	Negotiations	?

ZI overview - https://www.zi.uzh.ch/de/staff/software-elearning/ai-tools-und-services.html
ZI blog, KI category - https://www.uzh.ch/blog/zi/category/ki/
Übersicht zu KI-Tools und Services an der UZH (ZI blog, 15.06.2026) - https://www.uzh.ch/blog/zi/2026/06/15/uebersicht-zu-ki%E2%80%91tools-und-services-an-der-uzh/

Trying to move the ceiling…

The proposal: evaluate a self-hosted frontier open-weights model behind the ZI LLM Proxy.

e.g. Kimi K3: weights released 27.07.2026 · 2.8T MoE, 104B active · 1M context · 1.56 TB of weights · custom licence permitting research and internal deployment

Pilot workloads: literature assessment; qualitative coding of interview transcripts; …

Self-hosted means no data leaves the institution, which is exactly what the classification ceiling is protecting against.

Moonshot AI releases weights for Kimi K3 - https://www.tomshardware.com/tech-industry/artificial-intelligence/moonshot-ai-releases-weights-for-kimi-k3-firing-a-shot-across-the-bow-of-openai-and-anthropic-open-weight-model-performs-almost-as-well-as-frontier-models-while-being-2-3x-easier-to-run

Simon Willison, Kimi K3 notes (27.07.2026) - https://simonwillison.net/2026/Jul/27/kimi-k3/

Moonshot opens Kimi K3 weights under a revenue-tiered licence - https://www.unite.ai/moonshot-opens-kimi-k3-weights-under-a-revenue-tiered-license/

… and what stands in the way

Naming the obstacles honestly is part of the argument.

- **Hardware:** multi-node, and not cheap
- **Licence:** bespoke terms, needs a proper legal read
- **CISO review:** Open models are objects of foreign origin (KIMI = China). Self-hosting removes the data transfer, but not the question

Governance seen from the side of someone trying to extend a policy, rather than only apply it.

→ RoomPulse

Code 81197

LAB
Information Technology & Experimental Ethics

University of Zurich[UZH]
Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

Here the focus was on data protection, tools, (mis)match.

The October course walks across three days and a wider map.

What it adds:

- Research on AI vs research with AI - and what changes in design, validity and ethics
- Tools mapped by research task, not by data type
- Context management and writing with AI
- A workflow of your own

No preparatory work.

Rate me!

ITE LAB
Information Technology & Experimental Ethics

University of Zurich[UZH]
Institute of Biomedical Ethics and History of Medicine

Giovanni Spitale, PhD

Let’s keep talking

giovanni.spitale@ibme.uzh.ch

Download the slides

Information Technology & Experimental Ethics

University of Zurich[UZH]

Institute of Biomedical Ethics and History of Medicine