UNIVERSITÄT HOHENHEIM | Computational Science Hub

Innovating for Impact: AI Governance Education for Inclusive Societies

Dr. Johannes Bleher

October 1, 2025

Context

Course Design (AE4AI) (Weuts, Bleher, Bleher, Flores, Xuanyang, Pujszo and Almási 2025)

Ecosystem White Paper (ASEF) (Weuts, Billones, Bleher, Pujszo, Flores, Almási, Xuanyang, Soh, Rivera, Tozsa, Cachapero and Hammerbauer 2025)

- Failure taxonomy (2×2+1)
- AI law: EU vs. China
- Week-by-week outputs and ideas for a sandbox implementation

Why Now?

- AI is reshaping **economies**, **societies**, and **education** at unprecedented speed.
- AI is reshaping academic writing: Are you writing your papers *with* AI or already *for* AI?
- Governance challenge: move from *principles* to *operational practice* (law, policy, oversight).
- Universities can act as **laboratories for inclusive AI governance**.

Why Universities?

Universities can operationalize AI governance by coupling technical stacks (data, infrastructure) and teaching (methods, law, ethics) with governance practice (DPIAs, audits, sandbox rules).

- Not another ethics add-on
  Instead: **governance-by-design** in code, data, exams, and workflows.
- Measured by artifacts, behaviors, and external impact (*not* just course evaluations).

Five Operational Failures (At a Glance)

Representation

Biased / drifting data

- Historical bias $\rightarrow$ unfair recommendations
- Feedback loops (e.g. predictive policing)
- Fix: datasheets, sampling audits, drift monitors

Specification

Proxy targets / confounders

- "Clicks = satisfaction" $\rightarrow$ clickbait
- Confounding: Fever or Tylanol contributing to autism?
- Fix: causal checks, counterfactual evaluation

Five Operational Failures (At a Glance)

Generalization

OOD brittleness / Goodhart

- Benchmark gaming, reward hacking
- Weather/locale shift breaks model
- Fix: stress tests, red-teaming, eval sets by shift

Interaction

Human misuse / overtrust

- Automation bias with fluent LLMs
- Ambiguous affordances $\rightarrow$ wrong use
- Fix: usage bounds, UX warnings, literacy

Governance *Oversight and strategic failures*

- Hidden reasoning / deceptive behavior; weak auditability
- Multi-agent dynamics (collusion, races to the bottom)
- Fix: audit trails, DPIA, incident response, external review

International Regulation (Contrast)

European Union

- *One framework law:* EU AI Act + GDPR/DSA/DMA
- *Risk-based tiers:* Unacceptable, High-, Limited- and Minimal-risk; GPAI duties apply
- *Transparency duties:* logs, model/data cards, DPIAs
- *Governance:* centralized via EU AI Office + national regulators

China

- *Patchwork laws:* PIPL (privacy), DSL (security), sector rules
- *Focus areas:* Generative AI with "public opinion / mobilization" risks
- *Consent regime:* tiered ("general" vs. "separate" consent)
- *Governance:* decentralized, CAC emerging as coordinator

The Alignment Problem in the Three Arenas

Econometrics	**model** vs **true economic behavior** *Forecast failures, proxy misuse, distorted incentives*
Artificial Intelligence	**specified metric** vs **societal values** *Reward hacking, distribution shifts, misalignment*
Teaching	**exam** vs **graduate skills for real-world practice** *Students optimize for tests, not for competence*

From Data to Governance

- **ASEF Working Group (EU–Asia):** move beyond the siloed *B–A–G* (Build–Assess–Govern).
- Integrate **technical literacy** with **law/policy**, **ethics**, and **societal perspectives**.
- Translate principles into **actionable institutional practice**.

Universities as Governance Labs

What this looks like in practice

- Embedded ethics in core technical courses.
- Governance documentation: model cards, DPIAs, audit trails.
- Role-play: *regulator–industry–civil society* simulations.
- Community partnerships (NGOs, public agencies, industry).

Why universities?

- Neutral conveners across stakeholders.
- Safe-to-fail environments $\rightarrow$ *learned vigilance*.
- Talent pipelines with *governance fluency*.

Curriculum Architecture

Building Blocks

W1-2	AI & Society (foundations, cases)
W3-5	Technical Foundations (DL, RAG, robustness)
W6-8	Operational Failures (repr./spec./gen./interact./govern)
W9-10	Law & Regulation (EU AI Act, PIPL, etc.)
W11-12	Ethics (principles, justice, power)
W13-14	Interdisciplinary & Stakeholder Methods (VSD, embedded)
W15-16	Experiential Project (risk, compliance, policy)

Threaded Throughout

- Justice, equity, inclusion
- Community engagement
- Faculty development
- Assessment & evidence

Suggestions for Concrete Course Artifacts

- W1–2: Stakeholder map
- W3–5: Model card v1 + tests
- W6–8: Bias audit + drift monitor
- W9–10: Mini-DPIA (EU/China variants)
- W11–12: Decision log (trade-offs, harms)
- W13–14: Stakeholder interview memo
- W15–16: Audit packet (repo, docs, Continuous Integration)
- W17: Post-mortem + rubric self-score

Illustrative Cases

- **Unconventional data → policy:** search/activity data for inflation/consumption → bias & governance concerns.
- **HF markets:** order book dynamics → misalignment propagation, oversight needs.
- **Education analytics:** socio-educational data projects → fairness, consent, and DPIA practice.

Governance Sandbox

Stack as Sandbox

- JupyterHub + GitLab: reproducibility, logs, reviews.
- Controlled exams (SEB): *own skill* assessment.
- Templates: model cards, risk registers, audit checklists.

Learning Outcomes

- Diagnose failure modes; propose mitigations.
- Navigate regulation across jurisdictions.
- Engage stakeholders; justify trade-offs.

Call to Action

- **Integrate** ethics, law, and policy into technical teaching *by default*.
- **Co-produce** curricula with communities, industry, and policymakers.
- **Institutionalize** sandboxes: continuous update, audits, faculty training.

Connect. Equip. Achieve.

Connect

break silos • include different voices

Equip

methods + infrastructure + governance practice

Achieve

responsible AI for inclusive societies

References

Weuts, R., Billones, R. K., Bleher, J., Pujszo, P., Flores, R. T., Almási, Z., Xuanyang, G., Soh, J., Rivera, C., Tozsa, R., Cachapero, C. and Hammerbauer, M.: 2025, White Paper for Universities on Navigating Artificial Intelligence Innovation Ecosystems in the area of AI Governance. Accessed: 2025-03-15.

Weuts, R., Bleher, J., Bleher, H., Flores, R. T., Xuanyang, G., Pujszo, P. and Almási, Z.: 2025, Ai governance in higher education: A course design exploring regulatory, ethical and practical considerations. Submitted on 7 September 2025.