UNIVERSITY OF ZURICH • FACULTY OF LAW • YOAN HERMSTRÜWER

Law and Artificial Intelligence

Course Syllabus

Instructor	Yoan Hermstrüwer
Faculty	University of Zurich, Faculty of Law
Level	Advanced
Format	7 Modules · 14 Sessions · 90 minutes each
Language	English
Prerequisites	None — no advanced mathematics assumed; formal tools introduced from scratch
Assessment	Written examination, closed book

Course Description

What is a machine learning model, and where does the law attach to it? Why are neural networks "black boxes," and what does that mean for the right to an explanation? Can fairness be defined — and if competing definitions cannot all hold at once, who decides? When an AI system decides, what does meaningful human oversight actually require?

This course equips students to reason rigorously about artificial intelligence — not as metaphor or buzzword, but as a set of concrete technical systems with concrete legal consequences. Its central proposition is that sound legal reasoning about AI is impossible without a real understanding of how these systems work. The first half of the course therefore demystifies the technology — machine learning, neural networks, data processing, and large language models — at a genuine conceptual–technical level. No advanced mathematics is assumed; formal tools are introduced from scratch.

The second half turns to the conceptual problems that follow: accuracy, discrimination, fairness, explainability, automation and oversight, and value alignment. The emphasis throughout is on conceptual problems rather than on doctrine. Regulatory material (the EU AI Act, the GDPR, the Swiss and Council of Europe frameworks) is introduced where it attaches to a technical or conceptual issue, then consolidated near the end, so that regulation appears as the answer to problems students already understand.

Two running cases thread through the course, met first as technical objects and later as legal problems: **COMPAS** (recidivism risk scoring), which runs through the machine-learning, data, accuracy, discrimination, and fairness sessions and resurfaces in *State v. Loomis*; and **Mata v. Avianca** (the lawyer sanctioned for AI-fabricated citations), which anchors the session on large language models.

Learning Objectives

By the end of the course, students will be able to:

1. **Explain**, at a conceptual level, how machine learning, deep neural networks, and large language models work and how they are trained.

2. **Locate** where legal questions attach across the AI lifecycle — in the data, the labels, the model, the deployment, and the human oversight around it.
3. **Analyze** the core conceptual problems of AI and articulate why several resist purely technical solutions.
4. **Evaluate** the principal regulatory responses (EU AI Act, GDPR, Swiss and Council of Europe frameworks) as answers to those problems.
5. **Assess** real and proposed uses of AI in the legal system with both technical and normative literacy.

Assessment

Assessment is by a single **written examination** at the end of the course. The examination is **closed book**: no materials, notes, or electronic aids are permitted. There is no term paper, no take-home component, and no oral examination.

Written examination (closed book)	100%
Duration	120 minutes
Permitted aids	None

The exam covers the whole course and is designed to test **conceptual command rather than recall of bibliographic detail**. Students should expect two kinds of question: short questions asking them to explain a technical concept and its legal significance (for instance, what a loss function does and why its choice is a normative decision; why a deep network is opaque in a sense that disclosure alone cannot cure); and a longer problem question asking them to analyze a described AI system — classifying it under the relevant regulatory framework, locating where in its lifecycle the legal problems arise, and assessing the arguments on each side. No question requires mathematical derivation. The exam covers the sessions, and the **core and background readings** are useful in preparing for the exam.

A Note on the Readings

Each session lists readings in two tiers. **Core** readings are central: two to four items per session, chosen so that the weekly load is one that can be met with care. **Background** readings provide additional information: they deepen a point, supply technical detail, or document a debate, and several are reference works to be consulted selectively rather than read end to end.

Course Plan

Module I — Foundations of Artificial Intelligence *Sessions 1–2*

Session	Title	Topics
1	What is AI?	Symbolic, rule-based systems vs. learned, statistical systems; a compressed history from Dartmouth (1956) to the transformer (2017). Defining "AI system" as a legal act: the EU AI Act's definitional perimeter; the rule/standard tension mapped onto rule/learning.
2	Supervised and Unsupervised Learning	Learning a function from labeled examples; classification vs. regression; clustering; a sketch of reinforcement learning. Correlations, not understanding. Prediction vs. precedent; COMPAS enters as a supervised classifier.

Readings — Session 1

Core

- Surden, *Artificial Intelligence and Law: An Overview*, 35 Georgia State University Law Review 1305 (2019).
- Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 UC Davis Law Review 399 (2017).
- Russell & Norvig, *Artificial Intelligence: A Modern Approach* (4th ed., Pearson, 2021), ch. 1.

Background

- McCarthy, Minsky, Rochester & Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* (1955), reprinted in AI Magazine 27(4) (2006), 12–14.
- Nilsson, *The Quest for Artificial Intelligence* (Cambridge University Press, 2009), selected chapters.

Readings — Session 2

Core

- Surden, *Machine Learning and Law*, 89 Washington Law Review 87 (2014).
- Domingos, *A Few Useful Things to Know About Machine Learning*, Communications of the ACM 55(10) (2012), 78–87.
- Kleinberg, Ludwig, Mullainathan & Obermeyer, *Prediction Policy Problems*, American Economic Review 105(5) (2015), 491–495.

Background

- Alpaydin, *Machine Learning: The New AI* (MIT Press, 2016), selected chapters.
- Hermstrüwer & Imhof, *Cross-Jurisdictional Machine Predictions in Antitrust* (working paper, 2026).

Module II — Deep Learning and Data Processing *Sessions 3–4*

Session	Title	Topics
3	Neural Networks and Deep Learning	A neuron as a weighted sum plus a nonlinearity; depth as automatic feature learning; loss, gradient descent, back-propagation; overfitting vs. generalization. The technical origin of the black-box problem: no intent, no reasons — a minimized loss.
4	Data Processing and the Training Pipeline	The full lifecycle: collection, labeling, splitting, training, evaluation, deployment, monitoring. Data, not code, as the real source of an ML system; "ground truth" as a human and legal choice. Worked example: COMPAS's label (rearrest, not reoffense).

Readings — Session 3

Core

- LeCun, Bengio & Hinton, *Deep Learning*, Nature 521 (2015), 436–444.
- Nielsen, *Neural Networks and Deep Learning* (Determination Press, online, 2015), ch. 1–2.
- Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, Big Data & Society 3(1) (2016), 1–12.

Background

- Goodfellow, Bengio & Courville, *Deep Learning* (MIT Press, 2016), ch. 6.
- Krizhevsky, Sutskever & Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, NeurIPS 25 (2012).

Readings — Session 4

Core

- Lehr & Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 UC Davis Law Review 653 (2017).
- Gebru et al., *Datasheets for Datasets*, Communications of the ACM 64(12) (2021), 86–92.
- Angwin, Larson, Mattu & Kirchner, *Machine Bias*, ProPublica (23 May 2016).

Background

- Paullada, Raji, Bender, Denton & Hanna, *Data and its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research*, Patterns 2(11) (2021), 100336.
- Crawford, *Atlas of AI* (Yale University Press, 2021), ch. 3 ("Data").

Module III — Language and Machines *Sessions 5–6*

Session	Title	Topics
5	Foundations of NLP	Text as data and the representation problem: tokenization, bag-of-words to embeddings, meaning as geometry, attention at a conceptual level. NLP as the AI closest to legal work (search, e-discovery, contract review); bias baked into representations; legal-LLM evaluation as the research frontier.
6	Large Language Models and Generative AI	Pretraining as next-token prediction, fine-tuning, RLHF; in-context learning, retrieval-augmented generation, agents, and failure modes. Plausibility rather than verified truth; *Mata v. Avianca* pays off; RLHF plants the seed for value alignment.

Readings — Session 5

Core

- Jurafsky & Martin, *Speech and Language Processing* (3rd ed. draft), introductory chapter.
- Bolukbasi, Chang, Zou, Saligrama & Kalai, *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, NeurIPS 29 (2016).
- Fan, Ni, Merane, Tian, Hermstrüwer et al., *LEXam: Benchmarking Legal Reasoning on 340 Law Exams*, ICLR (2026) (arXiv:2505.12864).

Background

- Mikolov, Chen, Corrado & Dean, *Efficient Estimation of Word Representations in Vector Space*, ICLR Workshop (2013).
- Vaswani et al., *Attention Is All You Need*, NeurIPS 30 (2017).

Readings — Session 6

Core

- Bender, Gebru, McMillan-Major & Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, FAccT (2021), 610–623.
- *Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443 (S.D.N.Y. 2023).
- Engel, Hermstrüwer & Kruse, *Can Doctrine Harness Silicon Law Clerks?* (MPI for Behavioral Economics Discussion Paper, 2026).

Background

- Brown et al., *Language Models are Few-Shot Learners*, NeurIPS 33 (2020).
- Bommasani et al., *On the Opportunities and Risks of Foundation Models* (Stanford CRFM, 2021), selected sections.

Module IV — Accuracy and Discrimination *Sessions 7–8*

Session	Title	Topics
7	Accuracy	What "accuracy" means and why a single number misleads: error types, base rates, calibration, benchmark pitfalls. Bridges to evidentiary standards and burdens of proof: which errors a legal system tolerates is a normative choice.
8	Discrimination	How discrimination arises in ML and maps onto legal categories: disparate treatment vs. impact, protected attributes and proxies, bias inherited from historical data. ECHR Article 14, EU equality directives, Swiss equality guarantees; the ProPublica/Northpointe exchange.

Readings — Session 7

Core

- Kleinberg, Lakkaraju, Leskovec, Ludwig & Mullainathan, *Human Decisions and Machine Predictions*, Quarterly Journal of Economics 133(1) (2018), 237–293.
- Raji, Bender, Paullada, Denton & Hanna, *AI and the Everything in the Whole Wide World Benchmark*, NeurIPS Datasets and Benchmarks Track (2021).
- Dieterich, Mendoza & Brennan, *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity* (Northpointe research report, 2016).

Background

- Guo, Pleiss, Sun & Weinberger, *On Calibration of Modern Neural Networks*, ICML, PMLR 70 (2017), 1321–1330.
- Liang et al., *Holistic Evaluation of Language Models*, Transactions on Machine Learning Research (2023), selected sections.

Readings — Session 8

Core

- Barocas & Selbst, *Big Data's Disparate Impact*, 104 California Law Review 671 (2016).
- Wachter, Mittelstadt & Russell, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI*, Computer Law & Security Review 41 (2021), 105567.
- Obermeyer, Powers, Vogeli & Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, Science 366 (2019), 447–453.

Background

- Hellman, *Measuring Algorithmic Fairness*, 106 Virginia Law Review 811–866 (2020).
- Zuiderveen Borgesius, *Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence*, The International Journal of Human Rights 24(10) (2020), 1572–1593.

Module V — Fairness and Explainability *Sessions 9–10*

Session	Title	Topics
9	Fairness	The formalization of fairness and its limits: competing criteria and the impossibility results showing they cannot all hold at once. Fairness as a contested normative choice that mathematics can sharpen but not settle — forcing the law to decide what fairness requires.
10	Explainability	Interpretability vs. post-hoc explanation, the limits of each, and the "right to explanation." Explainability for whom, to what end; the GDPR debate grounded in the mechanism from Session 3 rather than treated as a slogan.

Readings — Session 9

Core

- Kleinberg, Mullainathan & Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 8th Innovations in Theoretical Computer Science Conference (ITCS), LIPIcs 67 (2017), 43:1–43:23.
- Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, Big Data 5(2) (2017), 153–163.
- Binns, *Fairness in Machine Learning: Lessons from Political Philosophy*, FAT*, PMLR 81 (2018), 149–159.

Background

- Corbett-Davies, Gaebler, Nilforoshan, Shroff & Goel, *The Measure and Mismeasure of Fairness*, Journal of Machine Learning Research 24 (2023), 1–117.
- Selbst, Boyd, Friedler, Venkatasubramanian & Vertesi, *Fairness and Abstraction in Sociotechnical Systems*, FAT* (2019), 59–68.

Readings — Session 10

Core

- Wachter, Mittelstadt & Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, International Data Privacy Law 7(2) (2017), 76–99.
- Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, 87 Fordham Law Review 1085 (2018).
- Wachter, Mittelstadt & Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 Harvard Journal of Law & Technology 841 (2018).

Background

- Lipton, *The Mythos of Model Interpretability*, Communications of the ACM 61(10) (2018), 36–43.
- Doshi-Velez & Kim, *Towards a Rigorous Science of Interpretable Machine Learning* (arXiv:1702.08608, 2017).

Module VI — Human Oversight and Value Alignment *Sessions 11–12*

Session	Title	Topics
11	Human Oversight and Automation	Delegation and control: what meaningful human oversight requires and when "human-in-the-loop" is a fiction; automation bias and oversight failure modes. ADM in administrative law; *SyRI* and the childcare-benefits scandal; judicial review of automated public-sector decisions.
12	Value Alignment and Normative Benchmarks	Getting AI systems to pursue the goals we intend: the specification problem, reward hacking, RLHF — and the step from technical alignment to alignment with legal and constitutional values. Whose values, by what legitimate process?

Readings — Session 11

Core

- Citron, *Technological Due Process*, 85 Washington University Law Review 1249 (2008).
- Crootof, Kaminski & Price, *Humans in the Loop*, 76 Vanderbilt Law Review 429–510 (2023).
- Green, *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, Computer Law & Security Review 45 (2022), 105681.

Background

- Zerilli, Knott, Maclaurin & Gavaghan, *Algorithmic Decision-Making and the Control Problem*, Minds and Machines 29 (2019), 555–578.
- Cobbe, *Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making*, Legal Studies 39(4) (2019), 636–655.

Readings — Session 12

Core

- Gabriel, *Artificial Intelligence, Values, and Alignment*, Minds and Machines 30 (2020), 411–437.
- Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, NeurIPS 35 (2022).
- Engel, Hermstrüwer & Kim, *Human Realignment: An Empirical Study of LLMs as Legal Decision-Aids in Moral Dilemmas*, Artificial Intelligence and Law (2026).

Background

- Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking, 2019), selected chapters.
- Amodei, Olah, Steinhardt, Christiano, Schulman & Mané, *Concrete Problems in AI Safety* (arXiv:1606.06565, 2016).

Module VII — AI Regulation and AI Judicial Systems *Sessions 13–14*

Session	Title	Topics
13	AI Regulation	Consolidating the regulatory threads: the EU AI Act's risk-based architecture and its interaction with the GDPR; the Swiss sector-specific approach; the Council of Europe Convention; the Brussels effect. Regulation as the answer to problems already understood.
14	AI Judicial Systems	AI turned on the legal system itself: predictive justice, risk assessment, automated and hybrid adjudication, legal tech in practice. *State v. Loomis* ties COMPAS back to due process; procedural fairness in human, hybrid, and robot courts.

Readings — Session 13

Core

- Regulation (EU) 2024/1689 (Artificial Intelligence Act), OJ L, 12 July 2024 — selected provisions.
- Veale & Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act*, Computer Law Review International 22(4) (2021), 97–112.
- Council of Europe, *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, CETS No. 225 (2024); Switzerland signed 27 March 2025.
- Swiss Federal Council, *AI regulatory-policy decision* (12 February 2025): ratify the CoE Convention, applying primarily to state actors; legislative amendments as sector-specific as possible, with cross-sectoral rules limited to fundamental-rights areas such as data protection; consultation draft due by the end of 2026. Read with the revised Federal Act on Data Protection (FADP, in force 2023).

Background

- Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 Southern California Law Review 1529 (2019).

Readings — Session 14

Core

- *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016).
- Re & Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 Stanford Technology Law Review 242–289 (2019).
- Chen, Hermstrüwer, Langenbach, Stremitzer & Tobia, *Mitigating the Judicial Human–AI Fairness Gap*, Journal of Legal Analysis 18(1) (2026), 207–242.

Background

- Aletras, Tsarapatsanis, Preoţiuc-Pietro & Lampos, *Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective*, PeerJ Computer Science 2:e93 (2016).
- Sourdin, *Judge v Robot? Artificial Intelligence and Judicial Decision-Making*, 41(4) UNSW Law Journal 1114–1133 (2018).