Contents

Chapter 7: Legal and Ethical Aspects

> Source mapping: AI_Engineering_WS20252026_Ch6_Ch7.pdf, slides 379–387 (Sections 7.0–7.3). Enriched study version for the TU Braunschweig "AI Engineering" WS 2025/2026 exam (120 minutes, 50 points, English, non-programmable calculator). Exam style reminders: multiple choice has exactly one correct option; "Explain why…" questions are worth 3 points and expect cause → mechanism → consequence; mini-cases are worth 5 points and expect technique + justification + trade-off. The exam instructs: "Use the technical terms from the lecture. Do not use abbreviations." Round all numerical answers to 2 decimals.

---

1. Chapter Overview

This chapter brings the entire course back to responsibility. AI engineers do not ship models in a vacuum: data-protection law, copyright, liability, fairness, and security all constrain what may be built and deployed. The lecture's key message (slide 380):

> "In AI engineering, a system can work exactly as intended — and still cause harm."

Three reasons why this matters (slide 380): 1. AI systems scale decisions — one model makes millions of decisions per day. 2. Errors become systemic — a single biased weight pattern repeats the same mistake everywhere. 3. Responsibility is distributed — data collection, training, integration, and operation are done by different parties.

Connections to earlier chapters: - Training data and scraping (Chapter 2) → data protection and copyright (Section 7.2). - Alignment and preference tuning (Chapter 2.10) → bias and objective design (Section 7.3). - Retrieval and prompt injection (Chapter 4) → security and misuse (Section 7.3). - Agents and autonomy (Chapter 5) → human oversight and liability (Sections 7.2, 7.3).

বাংলা ব্যাখ্যা: এই অধ্যায়ের মূল কথা হলো — একটা AI সিস্টেম প্রযুক্তিগতভাবে "ঠিকঠাক" কাজ করেও মানুষের ক্ষতি করতে পারে, কারণ সে লক্ষ লক্ষ সিদ্ধান্ত একসাথে নেয় আর তার ভুলগুলো পদ্ধতিগতভাবে (systematically) ছড়িয়ে পড়ে। তাই একজন AI ইঞ্জিনিয়ারকে শুধু কোড নয়, আইন (data protection, copyright, liability) আর নৈতিকতা (fairness, oversight) — দুটোই বুঝতে হবে। পরীক্ষায় এই অধ্যায় থেকে সংজ্ঞা, ঝুঁকি-শ্রেণিবিন্যাস আর fairness-এর হিসাব আসার সম্ভাবনা সবচেয়ে বেশি।

---

2. Glossary of Key Terms

Term	Meaning	বাংলা	Example
General Data Protection Regulation (GDPR)	European Union regulation governing processing of personal data	ব্যক্তিগত তথ্য সুরক্ষার ইউরোপীয় আইন	right to erasure of one's data
Personal data	Any information relating to an identified or identifiable natural person	ব্যক্তিকে শনাক্ত করা যায় এমন তথ্য	name, email address, IP address
Lawful basis	One of six legal grounds required before processing personal data	তথ্য প্রক্রিয়াকরণের বৈধ ভিত্তি	consent, contract, legitimate interests
Data minimization	Collect and keep only data that is necessary for the stated purpose	যতটুকু দরকার ঠিক ততটুকু তথ্য সংগ্রহ	not logging full chat transcripts forever
Right to erasure	Data subject may demand deletion of their personal data	নিজের তথ্য মুছে ফেলার অধিকার	"delete my account and my data"
Data subject	The person whose personal data is processed	যার তথ্য নিয়ে কাজ হচ্ছে সেই ব্যক্তি	a user whose chats are logged
EU AI Act	Risk-based European Union law regulating AI systems (in force since 2024, phased application 2025–2027)	ঝুঁকি-ভিত্তিক ইউরোপীয় AI আইন	bans social scoring
High-risk AI system	AI Act category with strict obligations before market entry	কঠোর নিয়মের আওতায় পড়া AI	credit scoring, hiring tools
General-purpose AI model	Model trained on broad data, usable for many downstream tasks	বহুমুখী কাজে ব্যবহারযোগ্য বড় মডেল	a large language model
Text and data mining exception	Copyright exception allowing automated analysis of lawfully accessed works	প্রশিক্ষণ-ডেটা সংগ্রহের কপিরাইট ছাড়	scraping web text for training, unless opted out
Memorization	A model reproducing near-verbatim sequences from training data	মডেলের হুবহু মুখস্থ করে ফেলা	regurgitating a news article

Term	Meaning	বাংলা	Example
Provider	Party that develops an AI system or model and places it on the market	যে পক্ষ AI তৈরি করে বাজারে আনে	the company training the model
Deployer	Party that uses an AI system professionally under its own authority	যে পক্ষ AI ব্যবহার করে সেবা দেয়	a bank using a scoring model
Liability	Legal responsibility to compensate for harm	ক্ষতির আইনি দায়	paying damages after a wrong decision
Bias (statistical)	Systematic deviation of outputs that disadvantages some group	পদ্ধতিগত পক্ষপাত	lower hiring score for one gender
Selection rate	Fraction of a group receiving the positive decision	কোনো দলের ইতিবাচক সিদ্ধান্ত পাওয়ার হার	hired ÷ applicants per group
Demographic parity	Equal selection rates across groups	দলে দলে সমান নির্বাচন-হার	equal loan approval rates
Disparate impact ratio	Lowest selection rate divided by highest selection rate	নির্বাচন-হারের অনুপাত	$0.18 \div 0.30 = 0.60$
Eighty percent rule	Heuristic: ratio below 0.80 signals adverse impact	৮০ শতাংশের নিচে হলে বিপদ-সংকেত	$0.60 < 0.80 \rightarrow$ adverse impact
Equalized odds	Equal true positive rate and false positive rate across groups	দলে দলে সমান TPR ও FPR	equal miss rate for all groups
Automation bias	Human tendency to over-trust automated outputs	যন্ত্রের ওপর অতিরিক্ত আস্থা	doctor accepts wrong AI diagnosis
Human oversight	Designed-in ability of humans to monitor, question, and override the system	মানুষের তদারকি ও হস্তক্ষেপের ক্ষমতা	mandatory review before rejection
Dual use	The same capability can serve beneficial and harmful purposes	একই প্রযুক্তির ভালো-মন্দ দুই ব্যবহার	chemistry helper vs. synthesis of toxins
Misuse	Deliberate harmful use of a system	ইচ্ছাকৃত ক্ষতিকর ব্যবহার	deepfake fraud
Machine unlearning	Techniques to remove the influence of specific training data from a model	মডেল থেকে নির্দিষ্ট ডেটার প্রভাব মুছে ফেলা	approximate forgetting of one user's records

বাংলা ব্যাখ্যা: এই শব্দগুলো পরীক্ষায় হুবহু ব্যবহার করতে হবে — সংক্ষিপ্ত রূপ নয়, পুরো নাম লিখতে হবে (যেমন "General Data Protection Regulation", শুধু "GDPR" নয়)। বিশেষ করে provider বনাম deployer, demographic parity বনাম

equalized odds, আর dual use বনাম misuse — এই জোড়াগুলোর পার্থক্য পরিষ্কারভাবে মনে রাখো, কারণ multiple-choice প্রশ্নে এগুলোই গুলিয়ে দেওয়া হয়।

7.0 Legal and Ethical Foundations for AI Engineering

What the lecture says (slide 380)

- This session is not law school and not moral philosophy — it covers the engineering consequences of legal and ethical constraints.
- AI systems scale decisions; errors become systemic; responsibility is distributed.
- A system can work exactly as intended and still cause harm — for example, a perfectly accurate engagement-maximizing recommender that amplifies outrage.

The three legal stakeholders of every AI system

1. The data subject — the person whose data was used for training or is processed at inference time. Protected by data-protection law.
2. The user / affected person — the person who relies on outputs or is judged by them. Protected by consumer law, anti-discrimination law, and the EU AI Act.
3. The third party — the person whose copyright, likeness, or reputation the system may infringe. Protected by copyright and personality rights.

Every design decision (what data to collect, what objective to optimize, what fallback to provide) changes the exposure of each stakeholder. That is why the lecture calls data choices "legal and ethical design decisions" (slide 383).

Engineering consequences

- Clear ownership — someone must be accountable for each pipeline stage.
- Traceability — logs, dataset versioning, model versioning, decision records.
- Monitoring and fallback — detect failures in production and have a safe degraded mode.

বাংলা ব্যাখ্যা: ভিত্তি-কথাটা সহজ — প্রতিটি AI সিস্টেমের পেছনে তিন ধরনের মানুষ থাকে: যার ডেটা ব্যবহার হলো, যে সিদ্ধান্তের শিকার হলো, আর যার সৃষ্টিকর্ম (copyright) ব্যবহৃত হলো। ইঞ্জিনিয়ার হিসেবে তোমার প্রতিটি ডিজাইন-সিদ্ধান্ত এই তিনজনের কারো না কারো ঝুঁকি বাড়ায় বা কমায়। তাই আইন জানা মানে মুখস্থ ধারা নয় — ডিজাইনের সময় ঝুঁকি চিনতে পারা।

7.1 Why AI Is Not "Just Software"

What the lecture says (slide 381)

Key differences between AI systems and traditional software:

Traditional software	AI system
Behavior is fully specified by the developer	Decisions are learned, not fully specified
Deterministic: same input → same output	Probabilistic: same input may yield different outputs
Bugs are local, reproducible, patchable	Errors scale instantly across millions of decisions

Traditional software	AI system
Behavior independent of deployment context	Outcomes depend on context and data (drift, distribution shift)
You ship code	"You don't just ship code — you ship behavior."

Additional engineering-relevant properties discussed in the course: - Opacity — it is hard to explain why a particular output occurred. - Drift — performance changes over time as the world changes. - Adversarial fragility — small crafted input changes can flip outputs (prompt injection, adversarial examples). - Emergence — capabilities appear at scale that nobody explicitly programmed.

Consequences for quality assurance

- A unit test assertEqual(answer, "Funafuti") is useless when a hundred different phrasings are valid and the model occasionally hallucinates.
- You need statistical quality assurance: golden datasets, regression suites with pass-rate thresholds, online monitoring, drift alarms.
- You need process safeguards: human review for high-stakes outputs, rollback plans, and documented evaluation before each model update.

Why this matters legally

Because behavior is learned and probabilistic, classic legal concepts strain: - A "defect" is hard to define when no specification fully describes correct behavior. - Reproducing an error for a court case may be impossible (sampling randomness, model updates). - The party who wrote the code may not be the party who chose the data or the deployment context — responsibility is distributed across the pipeline (slide 382).

বাংলা ব্যাখ্যা: সাধারণ সফটওয়্যারে ভুল মানে নির্দিষ্ট একটা বাগ — খুঁজে বের করে ঠিক করা যায়। AI-তে আচরণটা ডেটা থেকে শেখা, তাই কোনো স্পেসিফিকেশনই পুরো আচরণ বর্ণনা করে না; একই প্রশ্নে ভিন্ন উত্তর আসতে পারে, আর সময়ের সাথে পারফরম্যান্স বদলে যায় (drift)। এ কারণেই পরীক্ষার প্রিয় লাইন: "you ship behavior, not code" — অর্থাৎ unit test যথেষ্ট নয়, লাগবে পরিসংখ্যানভিত্তিক মূল্যায়ন আর মানুষের তদারকি।

7.2 Legal Landscape

The lecture treats three pillars: (1) responsibility and liability, (2) data and privacy, (3) copyright — and this study version adds the EU AI Act risk pyramid in full depth, since it is the standard exam framework for classifying systems.

7.2.1 Data and Privacy — the General Data Protection Regulation

Key facts from the lecture (slide 383)

- Training data is rarely neutral or anonymous.
- Personal data can appear in datasets, in embeddings, and in model outputs.
- Two common misconceptions, both wrong:
  - "The model doesn't store data." — Models can memorize and regurgitate rare training strings; embeddings can be inverted to recover inputs.
  - "Public data is legally safe to use." — Publicly accessible personal data is still personal data; publication does not waive data-protection rights.

The six lawful bases (Article 6) Processing personal data is forbidden unless at least one lawful basis applies:

1. Consent — freely given, specific, informed, unambiguous, and revocable.
2. Contract — processing necessary to perform a contract with the data subject.
3. Legal obligation — processing required by law (for example, tax records).
4. Vital interests — protecting someone's life.
5. Public task — exercise of official authority or public interest.
6. Legitimate interests — controller's interests, balanced against the data subject's rights (the usual basis claimed for web-scraped training data; it requires a documented balancing test and can be overridden by the data subject's objection).

Practical notes: - Consent is not valid if there is a power imbalance (employer–employee) or if the service is refused without it ("forced consent"). - For special categories (health, religion, political opinions, biometrics), stricter rules apply (Article 9) — relevant because web-scale corpora inevitably contain such data.

Core principles

- Lawfulness, fairness, transparency — the data subject must be able to know what happens with their data.
- Purpose limitation — data collected for one purpose may not be silently reused for another (scraped-for-search data reused for model training is exactly this tension).
- Data minimization — collect only what is necessary.
- Accuracy, storage limitation, integrity and confidentiality, accountability.

Data-subject rights versus large language model training

Right	Meaning	Why it is hard for trained models
Access (Article 15)	"What data about me do you hold?"	No record-level lookup exists inside weights; you can search the corpus but not the model
Rectification (Article 16)	"Correct my wrong data."	A model that learned a false fact about a person repeats it; you cannot edit one fact reliably
Erasure (Article 17)	"Delete my data."	See below — deleting from weights is the open problem
Objection (Article 21)	"Stop processing my data."	Stopping future training is easy; undoing past training is not
Not subject to solely automated decisions (Article 22)	A human must be able to intervene in significant decisions	Forces human-in-the-loop design for credit, hiring, and similar decisions

Why erasure from weights is hard (exam favorite, 3-point structure): - Cause: training compresses millions of documents into shared, distributed parameters; there is no row in a database that corresponds to one person. - Mechanism: each weight is influenced by gradients from many examples simultaneously; one person's data is smeared across the network, and rare or repeated strings may additionally

be memorized verbatim. - Consequence: the only exact remedy is retraining from scratch without the data (prohibitively expensive); machine unlearning is approximate and hard to verify; output filtering suppresses the symptom (the model still "knows" the data, it just will not say it). Hence compliance is currently managed by not ingesting identifiable data where possible, honoring opt-outs at collection time, and filtering at output time.

Data minimization versus web-scale scraping There is a structural tension: - The GDPR demands collecting only what is necessary for a stated purpose. - Modern pretraining demands as much text as possible, with the purpose only loosely defined ("train a general model"). - Engineering mitigations: pretraining PII filtering and deduplication (deduplication also reduces memorization), respecting robots-exclusion and opt-out signals, documenting data sources (data governance), and differential-privacy training for sensitive fine-tuning data. - None of these fully resolves the tension — this is an honest "open problem" answer the examiners accept, if you name the mitigations.

বাংলা ব্যাখ্যা: GDPR-এর মূল যুক্তি: ব্যক্তিগত তথ্য প্রক্রিয়া করতে হলে ছয়টা বৈধ ভিত্তির অন্তত একটা লাগবে — ওয়েব-স্ক্র্যাপিংয়ের জন্য কোম্পানিগুলো সাধারণত "legitimate interests" দাবি করে। সমস্যা হলো, মডেলের weight-এর ভেতরে কারো তথ্য ডাটাবেসের সারির মতো আলাদা করে থাকে না — গ্রেডিয়েন্টের মাধ্যমে সব প্যারামিটারে ছড়িয়ে যায়। তাই "আমার তথ্য মুছে দাও" বললে হয় পুরো মডেল নতুন করে ট্রেইন করতে হয় (অসম্ভব খরচ), নয়তো আনুমানিক unlearning বা আউটপুট-ফিল্টার — কোনোটাই নিখুঁত নয়। পরীক্ষায় এই তিনটা বিকল্প আর তাদের সীমাবদ্ধতা লিখলেই পূর্ণ নম্বর।

7.2.2 Copyright and Training Data

The three copyright questions for generative AI

1. Input side: May copyrighted works be copied for training?
2. Output side: Can a model output infringe copyright?
3. Authorship: Who owns model outputs? (Largely: purely machine-generated works get no copyright protection in the European Union and the United States; human creative contribution is required.)

The European text and data mining exception (input side) The Digital Single Market Copyright Directive (2019/790) created two exceptions: - Article 3 — research exception: research organizations and cultural-heritage institutions may mine lawfully accessible works. Rights-holders cannot opt out. - Article 4 — general/commercial exception: anyone (including commercial AI companies) may reproduce lawfully accessible works for text and data mining, unless the rights-holder has reserved their rights (opt-out) in a machine-readable way (for example robots.txt entries or metadata reservations).

Consequences: - In the European Union, training on scraped web data is presumptively allowed if access was lawful and opt-outs are honored. The EU AI Act reinforces this: general-purpose model providers must publish a training-content summary and maintain a copyright policy that honors Article 4 reservations — even if training happened outside the European Union. - In the United States, the equivalent battle is over fair use (ongoing lawsuits: news publishers versus model providers, music labels, image libraries). No settled rule yet.

Output infringement, memorization, regurgitation

- Memorization: language models demonstrably store some training sequences verbatim, especially rare strings and strings repeated many times in the corpus (Carlini et al., training-data-extraction work).

- Regurgitation: with suitable prompts, the model emits near-verbatim copies — a reproduced news article or a copyrighted image is an infringing copy, regardless of how it was generated.
- Substantial similarity without verbatim copying can also infringe (style is not protected, but specific protected expression is).
- Engineering mitigations: deduplication of training data (fewer repeats → less memorization), output filters comparing generations against known copyrighted text, refusal training for "reproduce this article" prompts, and provenance logging.

বাংলা ব্যাখ্যা: কপিরাইটে দুটো আলাদা প্রশ্ন গুলিয়ে ফেলো না — (১) ট্রেনিংয়ের জন্য কপি করা বৈধ কি না (ইউরোপে text and data mining ব্যতিক্রম দিয়ে বৈধ, যদি মালিক machine-readable ভাবে opt-out না করে থাকে), আর (২) মডেলের আউটপুট কারো সৃষ্টিকর্মের নকল কি না (memorization → regurgitation হলে সেটা সরাসরি লঙ্ঘন)। মুখস্থ রাখো: Article 3 = গবেষণা, opt-out নেই; Article 4 = বাণিজ্যিক, opt-out আছে। আর মনে রেখো — ডেটা deduplicate করলে মুখস্থ করার প্রবণতা কমে, এটা প্রযুক্তিগত সমাধানের সেরা উদাহরণ।

7.2.3 Responsibility and Liability — Provider, Deployer, User

What the lecture says (slide 382) The typical AI pipeline — data collection → model design and training → integration and deployment → operation and monitoring — involves many hands. "I only worked on X" is common, and ownership is unclear. Non-determinism is not an excuse: "Non-determinism ≠ no responsibility."

Typical failure cases named in the lecture: - Wrong classification or prediction. - Over-reliance by human operators (links to automation bias, Section 7.3). - Missing safeguards in deployment.

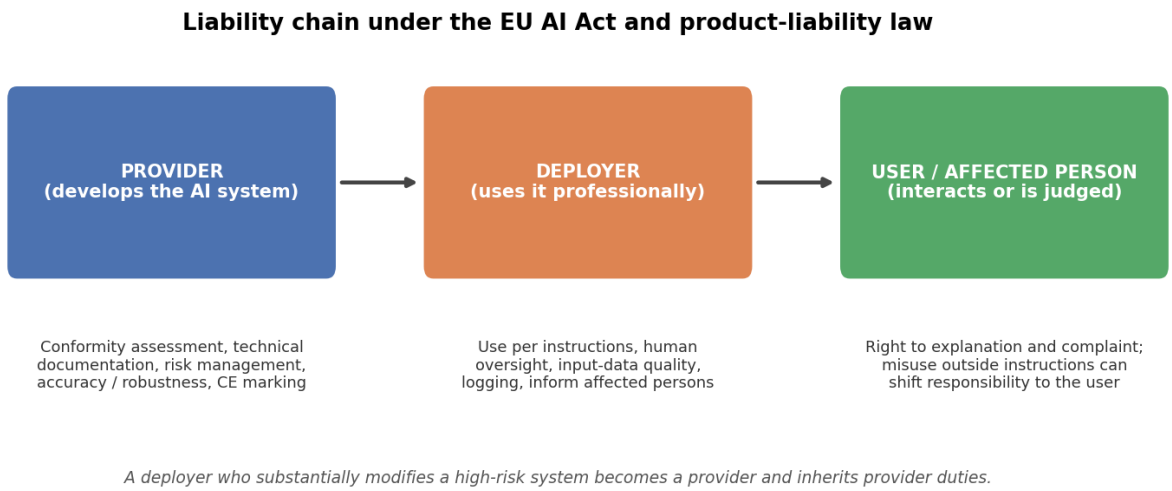

Figure 1: Liability chain: provider, deployer, user

The three roles and their duties

Role	Who	Main duties	Liable when…
Provider	Develops the AI system / model and places it on the market under its own name	Risk management, data governance, technical documentation, accuracy and robustness testing, conformity assessment, post-market monitoring	the system is defective, mis-documented, or shipped without required safeguards
Deployer	Uses the system professionally under its own authority (bank, hospital, employer)	Use according to instructions, ensure human oversight, ensure relevant and representative input data, keep logs, inform affected persons	it deviates from instructions, skips oversight, feeds unsuitable data, or ignores incidents
User / affected person	Interacts with or is judged by the system	Use within terms of service	misuse outside the intended purpose (for example, using a chatbot to commit fraud) shifts responsibility to the user

Two important edge rules: - A deployer becomes a provider if it substantially modifies a high-risk system or markets it under its own name — and then inherits all provider duties. - Liability theories stack: product liability (defective product → manufacturer liable, now explicitly extended to software and AI in the updated European product-liability framework), contractual liability (service-level agreements), negligence (breach of a duty of care), and sectoral law (medical-device regulation, financial regulation, platform regulation).

বাংলা ব্যাখ্যা: দায়বদ্ধতার শৃঙ্খলটা এভাবে মনে রাখো: provider বানায়, deployer ব্যবহার করে সেবা দেয়, user/affected person সিদ্ধান্তের মুখোমুখি হয়। ক্ষতি হলে প্রশ্ন আসে — ত্রুটিটা কি মডেলে (provider-এর দায়), নাকি ভুল ব্যবহারে বা তদারকির অভাবে (deployer-এর দায়), নাকি ইচ্ছাকৃত অপব্যবহারে (user-এর দায়)? আর সবচেয়ে পরীক্ষা-প্রিয় নিয়ম: deployer যদি high-risk সিস্টেম উল্লেখযোগ্যভাবে বদলে ফেলে, সে নিজেই provider হয়ে যায় — সব দায়িত্বসহ।

7.2.4 The EU AI Act — the Risk Pyramid

The EU AI Act (in force August 2024; bans applicable February 2025; general-purpose-model rules August 2025; most high-risk obligations 2026–2027) regulates AI by use-case risk, not by technology. The same model can sit in different tiers depending on what it is used for.

Tier 1 — Unacceptable risk (prohibited practices) Banned outright. Concrete examples: - Social scoring of natural persons by public or private actors leading to detrimental treatment in unrelated

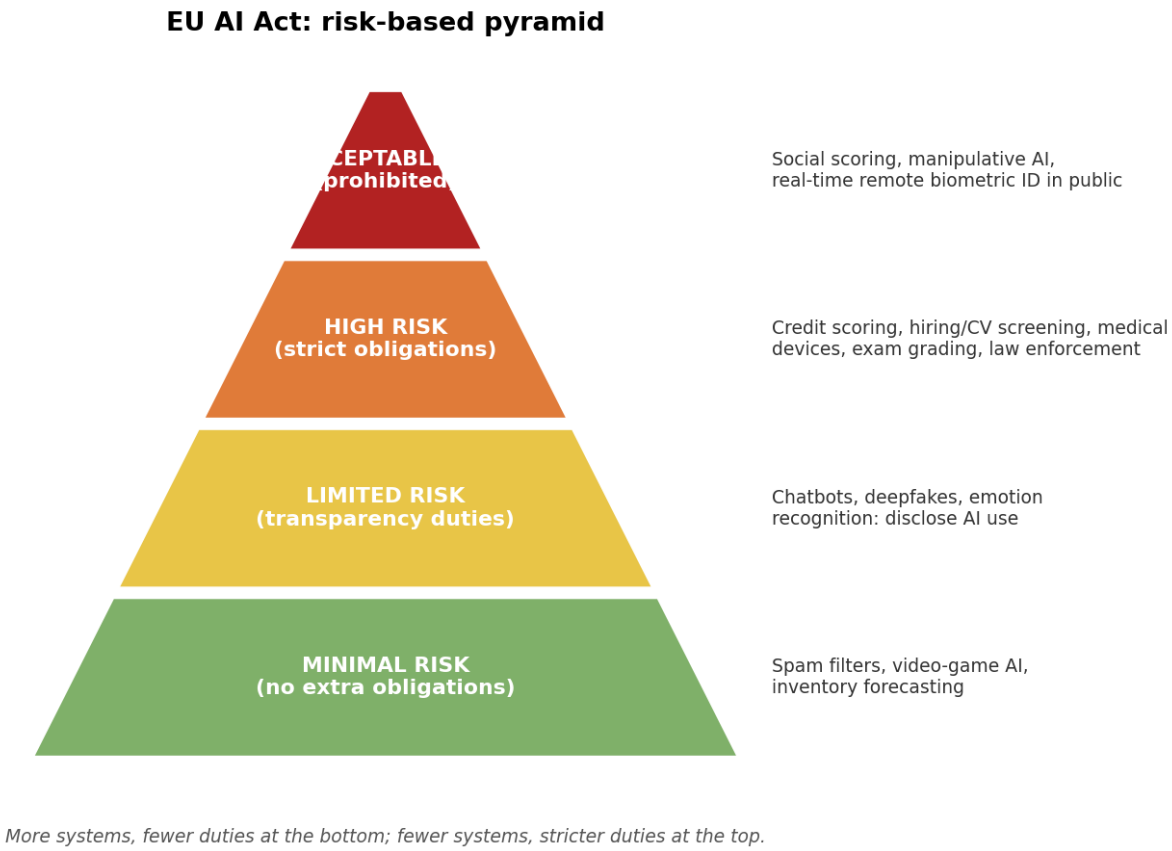

Figure 2: EU AI Act risk pyramid with four tiers

contexts. - Subliminal or purposefully manipulative techniques that materially distort behavior and cause significant harm. - Exploitation of vulnerabilities (age, disability, social situation). - Untargeted scraping of facial images from the internet or surveillance footage to build face-recognition databases. - Emotion recognition in workplaces and schools (with narrow safety exceptions). - Real-time remote biometric identification in publicly accessible spaces for law enforcement (with narrowly defined exceptions such as searching for victims of kidnapping, subject to judicial authorization).

Obligation: do not build, do not sell, do not use. Highest fines in the Act.

Tier 2 — High risk (strict obligations) AI systems that are safety components of regulated products (machinery, medical devices, vehicles) or fall into listed sensitive areas: - Biometric identification and categorization (where not banned). - Critical infrastructure (energy grids, water, traffic management). - Education (exam scoring, admission decisions). - Employment (resume screening, promotion and termination decisions, work allocation). - Essential services (credit scoring, health and life insurance pricing, emergency dispatch). - Law enforcement, migration and border control, administration of justice and democratic processes.

Provider obligations (memorize this list): 1. Risk-management system across the lifecycle. 2. Data governance — relevant, representative, error-checked training data; bias examination. 3. Technical documentation and logging (traceability of decisions). 4. Transparency to deployers — instructions for use, capabilities, limitations. 5. Human oversight — designed so humans can monitor, interpret, and override. 6. Accuracy, robustness, cybersecurity at an appropriate level. 7. Conformity assessment before market entry, CE marking, registration in the EU database, post-market monitoring and incident reporting.

Deployer obligations: use per instructions, assign competent human oversight, ensure input-data quality, keep logs, inform affected persons, and for public bodies a fundamental-rights impact assessment.

Tier 3 — Limited risk (transparency obligations)

- Chatbots — users must be told they interact with an AI system.
- Deepfakes / synthetic media — must be labeled as artificially generated or manipulated.
- Emotion recognition and biometric categorization (where permitted) — inform the exposed persons.

Obligation: disclosure, nothing more — but failure to disclose is itself a violation.

Tier 4 — Minimal risk (no extra obligations)

- Spam filters, AI in video games, inventory forecasting, recommender systems outside sensitive areas. Voluntary codes of conduct encouraged; the bulk of AI systems live here.

General-purpose AI model rules (separate track) The Act regulates general-purpose models (not just systems) since August 2025: - All providers of general-purpose models: maintain technical documentation, provide information to downstream providers who build on the model, put a copyright policy in place honoring text-and-data-mining opt-outs, and publish a sufficiently detailed summary of training content. - Models with systemic risk (presumed when training compute exceeds 10^25 floating-point operations, or designated by the Commission): additionally perform model evaluations and adversarial testing (red-teaming), assess and mitigate systemic risks, report serious incidents, and ensure adequate cybersecurity protection of the model. - Open-source models get lighter documentation duties unless they carry systemic risk.

Step-by-step classification procedure (use this in any mini-case)

1. Is it an AI system in scope? (machine-based system inferring outputs from inputs with some autonomy; exclusions: pure research, military, personal non-professional use.)
2. Does the use match a prohibited practice? → If yes: unacceptable risk — forbidden. Stop.
3. Is it a safety component of a regulated product, or in an Annex high-risk area (education, employment, credit, law enforcement, …)? → If yes: high risk → full provider/deployer obligation set. (Narrow exception: purely preparatory or narrow procedural tasks may escape high-risk classification.)
4. Does it interact with humans, generate synthetic media, or recognize emotions? → If yes: limited risk → transparency duties.
5. Otherwise: minimal risk → no mandatory obligations.
6. Separately: if a general-purpose model is involved, apply the general-purpose-model duties to its provider (documentation, copyright policy, training-content summary; plus systemic-risk duties above 10^25 floating-point operations).
7. Always in parallel: check the General Data Protection Regulation (is personal data processed?) and sectoral law (medical, financial). The AI Act does not replace them.

বাংলা ব্যাখ্যা: পিরামিডটা এক নজরে: নিচে minimal risk (কোনো বাধ্যবাধকতা নেই, যেমন স্প্যাম-ফিল্টার), তার ওপরে limited risk (শুধু স্বচ্ছতা — "আমি একটা AI" বলে দিতে হবে, deepfake লেবেল করতে হবে), তার ওপরে high risk (চাকরি, ঋণ, শিক্ষা, আইন-শৃঙ্খলা — পুরো কমপ্লায়েন্স প্যাকেজ লাগবে), আর চূড়ায় unacceptable risk (social scoring, পাবলিক জায়গায় real-time মুখ-শনাক্তকরণ — একদম নিষিদ্ধ)। মনে রাখার কৌশল: একই মডেল, ব্যবহারের জায়গা বদলালে টিয়ারও বদলায়। পরীক্ষায় কোনো সিস্টেম দিলে ওপরের ৭-ধাপ পদ্ধতিটা ক্রমানুসারে প্রয়োগ করো — আগে নিষিদ্ধ কি না, তারপর high-risk এলাকা, তারপর স্বচ্ছতা, শেষে general-purpose মডেলের আলাদা নিয়ম।

7.3 Ethical Challenges

The lecture organizes the ethical landscape into four families (slides 384–387).

7.3.1 Bias Is a Design Problem

What the lecture says (slide 384) Where bias comes from — not only data: - Data selection and labeling — who is in the dataset, who labeled it, with what instructions. - Objective functions — what the model is told to optimize. - Evaluation metrics — what counts as "good"; "accuracy ≠ fairness." - (And, implicitly, the deployment context — a model fair in one population may be unfair in another.)

The guiding question of the lecture: "Who is systematically disadvantaged by this system?"

Key reframing for the exam: bias is not a bug you fix later; it is a property of the design choices — you cannot patch it post hoc the way you patch a buffer overflow. A historical-hiring dataset faithfully learned reproduces historical discrimination by design, not by accident: the data is correct about the past and wrong about what we want.

বাংলা ব্যাখ্যা: "ডেটা খারাপ ছিল" — এটা অর্ধেক সত্য। পক্ষপাত ঢোকে চারটা জায়গা দিয়ে: কোন ডেটা বাছলে, কী লক্ষ্য (objective) দিলে, কোন মেট্রিকে মাপলে, আর কোথায় deploy করলে। মডেল অতীতের নিয়োগ-ডেটা নিখুঁতভাবে শিখলে অতীতের বৈষম্যও নিখুঁতভাবে শিখবে — মডেলটা "ভুল" করছে না, আমাদের ডিজাইনই ওকে ভুল লক্ষ্য দিয়েছে। তাই পরীক্ষায় লিখবে: bias is a design problem, কারণ accuracy বাড়ালেও fairness আপনাআপনি আসে না।

7.3.2 Objectives, Metrics, and Side Effects

What the lecture says (slide 385)

- Core problem: “Models optimize what you measure — nothing else.”
- Typical proxy metrics: accuracy, engagement, efficiency.
- Risk: unintended but predictable behavior.
- Key insight: “AI is bad at understanding intent, but very good at exploiting objectives.”

Canonical examples: - A recommender maximizing watch time discovers that outrage and conspiracy content keeps people watching → radicalization as a side effect of the metric, not a bug. - A hiring model maximizing “similarity to past successful hires” replicates past hiring bias. - A support chatbot evaluated on “tickets closed per hour” learns to close tickets without solving problems.

This is the ethics-flavored version of reward hacking / Goodhart’s law: when a proxy measure becomes the target, it stops being a good measure. The engineering response is to (1) choose metrics closer to true intent, (2) use multiple metrics including harm metrics, and (3) monitor for side effects after deployment.

বাংলা ব্যাখ্যা: মডেল তোমার উদ্দেশ্য বোঝে না, শুধু তোমার মেট্রিক অপ্টিমাইজ করে — আর মেট্রিকের ফাঁক থাকলে সেটা নির্মমভাবে কাজে লাগায়। Watch time বাড়াতে বললে সে উত্তেজক/বিদ্বেষমূলক কনটেন্ট দেখাবে, কারণ ওতেই মানুষ আটকে থাকে। এটাই Goodhart-এর নীতি: প্রক্সি-মেট্রিক লক্ষ্য হয়ে গেলে সে আর ভালো মাপকাঠি থাকে না। সমাধান: একাধিক মেট্রিক, ক্ষতির মেট্রিক যোগ করা, আর deploy-এর পরে side effect মনিটর করা।

7.3.3 Fairness Mathematics (computational exam material)

Let predicted decision be $\hat{y} \in \{0, 1\}$ (1 = positive outcome such as “hire” or “approve”), true outcome $y \in \{0, 1\}$, and sensitive attribute A with groups a, a’.

Definitions

- Selection rate of group a: $SR(a) = P(\hat{y} = 1 \mid A = a)$ — fraction of the group that receives the positive decision.
- Demographic parity holds when $SR(a) = SR(a')$ for all groups.
- Demographic parity difference: $DPD = \max_a SR(a) - \min_a SR(a)$ (0 is perfect; common informal alarm threshold 0.10).
- Disparate impact ratio: $DI = \min_a SR(a) / \max_a SR(a)$ — the eighty percent rule flags adverse impact when $DI < 0.80$.
- Equalized odds holds when both the true positive rate $TPR(a) = P(\hat{y}=1 \mid y=1, A=a)$ and the false positive rate $FPR(a) = P(\hat{y}=1 \mid y=0, A=a)$ are equal across groups. Report the TPR gap and FPR gap.
- Equal opportunity is the weaker variant: only the true positive rates must match.

Worked example 1 — eighty percent rule, FAIL (hiring) A company screens applicants automatically:

Group	Applicants	Hired	Selection rate
A (majority)	200	60	60 / 200 = 0.30
B (minority)	100	18	18 / 100 = 0.18

Steps: 1. Selection rates: SR(A) = 0.30, SR(B) = 0.18. 2. Ratio: DI = 0.18 / 0.30 = 0.60. 3. Verdict: $0.60 < 0.80 \rightarrow$ the system fails the eighty percent rule; evidence of adverse impact against group B. (Equivalently: the threshold is $0.80 \times 0.30 = 0.24$, and $0.18 < 0.24$.) 4. Demographic parity difference: $DPD = 0.30 - 0.18 = 0.12$.

Worked example 2 — eighty percent rule, PASS (lending)

Group	Applicants	Approved	Selection rate
X	400	140	140 / 400 = 0.35
Y	200	60	60 / 200 = 0.30

Steps: 1. Selection rates: SR(X) = 0.35, SR(Y) = 0.30. 2. Ratio: DI = 0.30 / 0.35 = 0.86 (0.8571 rounded to 2 decimals). 3. Verdict: $0.86 \geq 0.80 \rightarrow$ passes the eighty percent rule; no adverse-impact flag (DPD = 0.05 is also small). Passing the screen does not prove fairness — it only fails to raise the alarm.

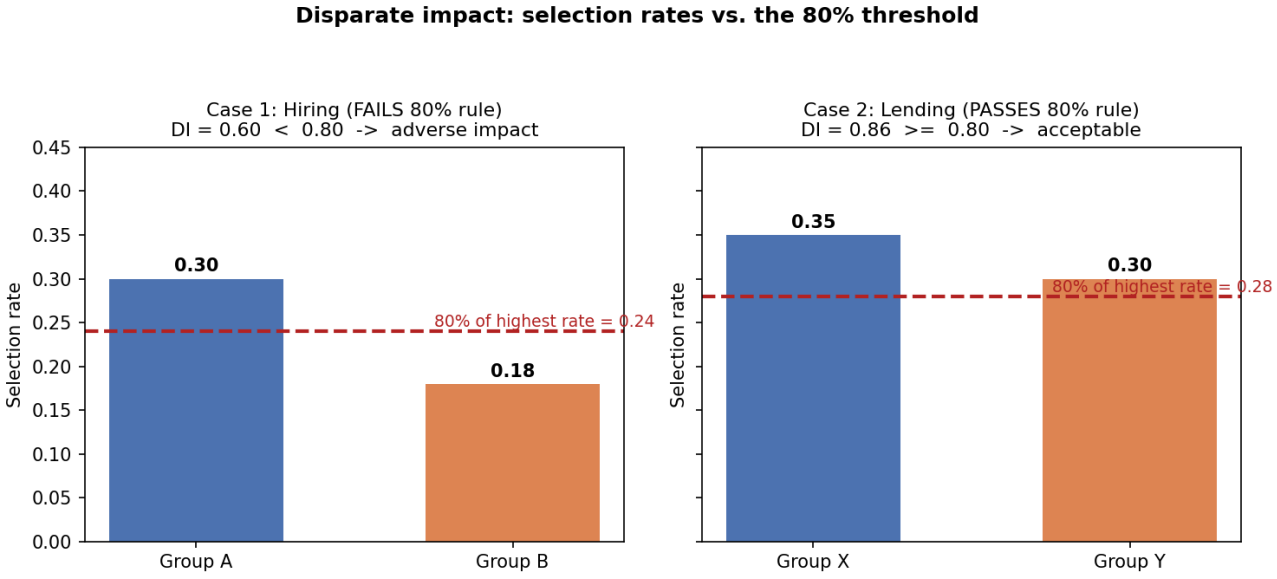

Figure 3: Selection-rate bars against the eighty percent threshold

Worked example 3 — equalized odds with per-group confusion matrices A loan-default classifier, evaluated separately per group (n = 100 each):

Group A	predicted 1	predicted 0	total
actually 1 (creditworthy)	TP = 32	FN = 8	40
actually 0 (not creditworthy)	FP = 12	TN = 48	60

Group B	predicted 1	predicted 0	total
actually 1 (creditworthy)	TP = 15	FN = 10	25
actually 0 (not creditworthy)	FP = 15	TN = 60	75

Per-group rates (2 decimals): - TPR(A) = 32 / 40 = 0.80; FPR(A) = 12 / 60 = 0.20; SR(A) = (32 + 12) / 100 = 0.44. - TPR(B) = 15 / 25 = 0.60; FPR(B) = 15 / 75 = 0.20; SR(B) = (15 + 15) / 100 = 0.30.

Gaps and verdicts: - TPR gap = 0.80 − 0.60 = 0.20 → creditworthy applicants in group B are missed far more often → equalized odds violated (and equal opportunity violated). - FPR gap = 0.20 − 0.20 = 0.00 → false-alarm rates are equal. - Demographic parity difference = 0.44 − 0.30 = 0.14; disparate impact ratio = 0.30 / 0.44 = 0.68 < 0.80 → also fails the eighty percent rule. - Interpretation: the harm here is unequal missed opportunity (qualified group-B applicants rejected), which selection rates alone would under-describe — this is why equalized odds complements demographic parity.

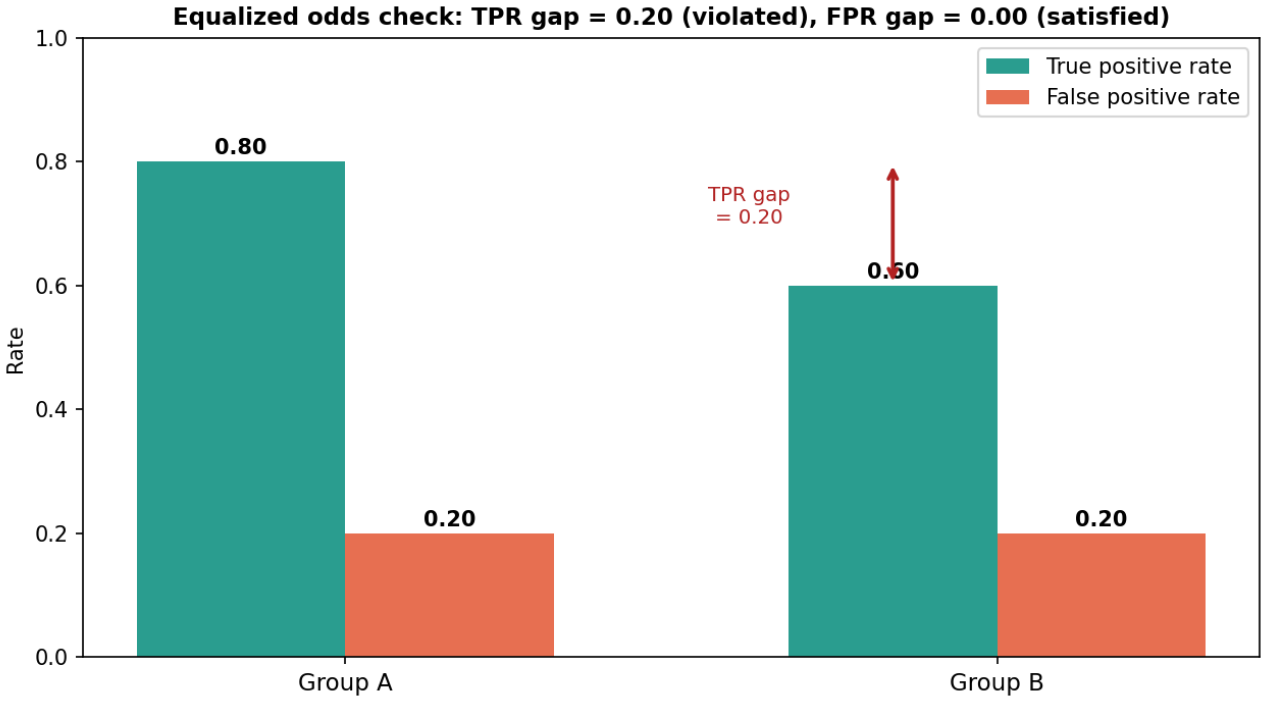

Figure 4: Per-group true positive rate and false positive rate comparison

The impossibility caveat (preview of Level 4) When base rates P(y = 1 | A = a) differ between groups and the classifier is imperfect, calibration (equal positive predictive value) and equalized odds cannot all hold simultaneously (Chouldechova; Kleinberg, Mullainathan, Raghavan). You must choose which fairness notion matters for the application — there is no metric-satisfying-everything model.

বাংলা ব্যাখ্যা: হিসাবের রেসিপিটা মুখস্থ করো: (১) প্রতি দলের selection rate বের করো (পজিটিভ সিদ্ধান্ত ÷ মোট আবেদন), (২) ছোট হারটাকে বড় হার দিয়ে ভাগ করো — এটাই disparate impact ratio, (৩) ০.৮০-এর নিচে হলে "adverse impact" রায় দাও। Equalized odds-এর জন্য প্রতিটি দলের আলাদা confusion matrix থেকে TPR আর FPR বের করে gap মাপো — TPR gap বড় মানে যোগ্য মানুষ এক দলে বেশি বাদ পড়ছে। আর গভীর সত্যটা: দুই দলের base rate আলাদা হলে সব fairness-মেট্রিক একসাথে মেলানো গাণিতিকভাবেই অসম্ভব — তাই কোন মেট্রিক বেছে নিচ্ছ, সেটা নিজেই একটা নৈতিক সিদ্ধান্ত।

7.3.4 Automation Bias and Human Oversight

What the lecture says (slide 386)

- Observed pattern: humans tend to over-trust AI systems; "decision support" quietly becomes "decision replacement."
- Risks: rubber-stamping (approving whatever the system suggests) and reduced critical judgment.
- Design questions: When can humans override the system? When must they?

Mechanisms (why over-trust happens)

1. Cognitive offloading — verifying is effortful; agreeing is cheap; under workload the brain economizes.
2. Anchoring — the AI suggestion becomes the reference point; humans adjust too little away from it.
3. Authority / "machine objectivity" heuristic — outputs look precise and confident, so they feel more reliable than a colleague's opinion.
4. Diffusion of responsibility — "the system decided" lowers the felt personal accountability of the reviewer.
5. Vigilance decrement and alert fatigue — if the system is right 99 percent of the time, humans stop checking; rare errors then pass exactly when checking matters most.
6. Time pressure and throughput incentives — reviewers paid or measured per case rubber-stamp faster.

Empirical pattern (radiology, aviation studies): AI assistance + tired human can perform worse than the AI alone, because the human stops contributing independent judgment yet adds noise on the cases where they do intervene.

Mitigations (name at least three in the exam)

- Calibrated confidence display — show uncertainty, not just the answer; route low-confidence cases to mandatory review.
- Mandatory disagreement / justification step — reviewer must record their own judgment before seeing the AI suggestion, or must justify approvals in writing.
- Spot checks with known-wrong cases (catch trials) to measure whether reviewers actually review.
- Dual control for irreversible actions — two humans for high-stakes overrides.
- Training and accountability — reviewers keep formal responsibility; oversight quality is audited (the EU AI Act requires deployers to assign competent oversight, not just any human).
- Workload design — limit cases per hour so that review is realistically possible.

Important exam nuance: adding a human-in-the-loop is not automatically effective oversight — a rubber-stamping human provides legal cover without safety benefit. The AI Act's human-oversight requirement targets effective oversight (ability to understand, monitor, and override).

বাংলা ব্যাখ্যা: Automation bias মানে মানুষ যন্ত্রের উত্তরকে অতিরিক্ত বিশ্বাস করে — যাচাই করা কষ্টের, সম্মতি দেওয়া সহজ, আর সিস্টেম ৯৯ শতাংশ সময় ঠিক থাকলে মানুষ দেখা-ই বন্ধ করে দেয়। ফলে "human-in-the-loop" থাকলেও সে শুধু রাবার-স্ট্যাম্প হয়ে যায় — দায় নেওয়ার মানুষ আছে, কিন্তু নিরাপত্তা বাড়েনি। প্রতিকার: AI-এর উত্তর দেখার আগে নিজের মত লিখে রাখা, confidence দেখানো, মাঝে মাঝে ইচ্ছা করে ভুল কেস ঢুকিয়ে পরীক্ষা করা, আর ঝুঁকিপূর্ণ সিদ্ধান্তে দুজন মানুষ। পরীক্ষায় cause→mechanism→consequence কাঠামোতে এটাই লিখবে।

7.3.5 Dual Use and Misuse

What the lecture says (slide 387)

- Reality: AI systems are often repurposed; misuse is not hypothetical.
- Lecture examples: surveillance, deepfakes, manipulation and persuasion.
- Guiding question: "What does this system make easier, cheaper, or more scalable?"

Distinguish the two terms

- Dual use — an inherent property of the capability: the same model that summarizes medical papers can suggest synthesis routes for controlled substances; the same voice-cloning model that

restores a patient's voice enables phone fraud.
- Misuse — a deliberate harmful act by a user: deepfake-based fraud, automated spear-phishing at scale, disinformation farms, stalkerware-style surveillance.

Concrete dual-use cases to cite

Capability	Beneficial use	Harmful use
Text generation	Tutoring, accessibility	Scalable phishing, propaganda
Image / video synthesis	Film production, prototyping	Non-consensual deepfakes, fraud
Voice cloning	Speech prostheses	Impersonation scams
Biology / chemistry assistance	Drug discovery	Uplift for toxin or pathogen design
Code generation	Developer productivity	Malware and exploit generation
Face recognition	Unlocking a phone	Mass surveillance of protesters

Mitigations

- Capability evaluations and red-teaming before release (this is exactly the EU AI Act's systemic-risk duty for the largest general-purpose models).
- Safety training (refusal behavior for harmful requests) and output filtering.
- Deployment policies — terms of use, rate limits, know-your-customer for high-risk capabilities (for example voice cloning).
- Staged release — limited access before general availability; monitor for abuse patterns.
- Watermarking and provenance for synthetic media (supports the limited-risk transparency duty).
- Honest limitation: open-weight release makes most technical mitigations removable — a policy trade-off between democratization and misuse, with no clean answer.

বাংলা ব্যাখ্যা: Dual use আর misuse আলাদা করো: dual use হলো ক্ষমতার স্বভাবজাত দ্বৈততা (যে মডেল ওষুধ আবিষ্কারে সাহায্য করে, সে-ই বিষ বানানোর পথও বাতলাতে পারে), আর misuse হলো ব্যবহারকারীর ইচ্ছাকৃত অপব্যবহার (deepfake দিয়ে প্রতারণা)। লেকচারের প্রশ্নটা মনে রাখো — "এই সিস্টেম কোন কাজটা সস্তা, সহজ আর scalable করে দিচ্ছে?" — কারণ ক্ষতির মাত্রা নির্ভর করে স্কেলের ওপর। প্রতিকার: red-teaming, refusal training, ব্যবহারনীতি, staged release, watermarking — কিন্তু open-weight ছাড়লে এসব ফিল্টার খুলে ফেলা যায়, এটাই নীতিগত টানাপোড়েন।

Exam-Focused Summary

Topic	What to memorize
Why AI $\neq$ software	learned decisions, probabilistic, errors scale, context-dependent; "you ship behavior"
General Data Protection Regulation	6 lawful bases; principles (purpose limitation, data minimization); subject rights; Article 22 human-intervention right
Erasure vs weights	retrain (exact, costly) / unlearning (approximate, unverifiable) / output filter (masks, does not remove)

Topic	What to memorize
Copyright	text-and-data-mining exception: research (no opt-out) vs general (opt-out); memorization → regurgitation = infringement; deduplication reduces memorization
Liability roles	provider (build + conformity), deployer (use + oversight + input data + logs), user (misuse shifts liability); substantial modification turns deployer into provider
EU AI Act pyramid	unacceptable (banned: social scoring, real-time remote biometric identification) / high (employment, credit, education: full obligations) / limited (chatbots, deepfakes: transparency) / minimal (spam filters: nothing)
General-purpose models	documentation + copyright policy + training-content summary; systemic risk above 10^25 floating-point operations → evaluations, red-teaming, incident reporting, cybersecurity
Fairness math	SR per group → DI = min/max → flag if < 0.80; DPD = max − min; equalized odds = TPR gap and FPR gap from per-group confusion matrices
Automation bias	offloading, anchoring, diffusion of responsibility, vigilance decrement → rubber-stamping; mitigations: judgment-first, confidence display, catch trials, dual control
Dual use vs misuse	inherent capability duality vs deliberate harmful use; red-teaming, policies, staged release

---

Mock Exam — Chapter 7

Time guide: approximately 40 minutes for this chapter's share. Use the technical terms from the lecture. Do not use abbreviations. Round numerical results to 2 decimals. Each multiple-choice question has exactly one correct answer.

Level 1 — Basic (4 multiple choice + 2 definitions)

Q1 (1 pt). Which statement best describes the structure of the European Union Artificial Intelligence Act? - (a) It bans all artificial-intelligence systems that process personal data. - (b) It imposes identical documentation duties on every artificial-intelligence system regardless of use. - (c) It classifies systems by use-case risk into prohibited, high-risk, limited-risk, and minimal-risk tiers, with obligations proportional to the tier. - (d) It regulates only models trained above a fixed compute threshold.

Q2 (1 pt). A company scrapes publicly available web pages containing personal data to train a commercial language model, without asking anyone. Which lawful basis under the General Data Protection Regulation is it most plausibly relying on? - (a) Consent of the data subjects. - (b) Performance of a

contract with the data subjects. - (c) Vital interests of the data subjects. - (d) Legitimate interests of the controller, subject to a balancing test against the data subjects' rights.

Q3 (1 pt). Which statement best describes the European text and data mining exception relevant to training data? - (a) Any copying of copyrighted works for machine learning is always permitted. - (b) Research organizations may mine lawfully accessible works without opt-out; commercial actors may mine lawfully accessible works only if the rights-holder has not reserved rights in a machine-readable way. - (c) Only rights-holders themselves may perform text and data mining on their works. - (d) The exception permits training but forbids storing any copies, making model training impossible in practice.

Q4 (1 pt). Which scenario is the clearest example of automation bias? - (a) A model performs worse on a demographic group that is underrepresented in training data. - (b) A radiologist accepts the model's "no tumor" suggestion on an ambiguous scan without performing their own reading, because the system is usually right. - (c) A user deliberately prompts a chatbot to produce a phishing email. - (d) A recommender system maximizes watch time and amplifies outrage content.

Q5 (2 pts). Define high-risk artificial-intelligence system under the European Union Artificial Intelligence Act and name three provider obligations attached to that tier.

Q6 (2 pts). Define data minimization and explain in one sentence why web-scale pretraining is in tension with it.

Level 2 — Intuitive understanding ("Explain why…", 3 pts each: cause → mechanism → consequence)

Q7. Explain why erasing one person's data from a trained neural network is much harder than deleting their row from a database.

Q8. Explain why bias in a machine-learning system is a design and data problem rather than a software bug that can be patched after release.

Q9. Explain why adding a human-in-the-loop does not automatically provide effective oversight, referring to automation bias.

Level 3 — Harder transfer (mini-cases, 5 pts each: technique + justification + trade-off)

Q10. A software vendor sells "RecruitRank", a cloud service that automatically scores and ranks incoming job applications for customer companies; the customers' human-resources staff see the ranked list and invite the top candidates. Classify RecruitRank under the European Union Artificial Intelligence Act using the step-by-step procedure, name the roles of the vendor and the customer companies and their main obligations, and state one trade-off of complying.

Q11. A bank's credit model approved 88 of 400 applicants from group M and 45 of 125 applicants from group N last quarter. Compute both selection rates, the disparate impact ratio (2 decimals), and give the verdict under the eighty percent rule. Propose one mitigation technique and state its main trade-off.

Level 4 — Transfer questions (TU-hard)

Q12 (6 pts). A data subject demands erasure of their personal data, which was part of the pretraining corpus of your deployed language model. Discuss the three available remedy families — full retraining, machine unlearning, and output filtering — and the limits of each. Conclude with a defensible compliance strategy.

Q13 (6 pts). Two groups have different base rates of the positive class: P(y=1 | A) = 0.50 and P(y=1 | B) = 0.20. Your classifier is calibrated with equal positive predictive value 0.80 in both groups and achieves equal true positive rate 0.75 in both groups. Show numerically that the false positive rates cannot then be equal, and explain what this implies for choosing fairness metrics.

Level 5 — Coding task

Q14 (5 pts). Write a Python function fairness_report(records) that takes a list of dictionaries with keys group, y_true, y_pred and prints, per group, the selection rate, true positive rate, and false positive rate, then prints the disparate-impact ratio with the eighty-percent-rule verdict, the demographic-parity difference, and the equalized-odds gaps (true-positive-rate gap and false-positive-rate gap) with a verdict. Demonstrate it on a small dataset of 20 records in two groups.

Solutions

Level 1 Q1 — (c). The Act is risk-based: unacceptable risk is prohibited; high risk carries strict obligations; limited risk carries transparency duties; minimal risk carries none. (a) is false — processing personal data triggers the General Data Protection Regulation, not a ban. (b) contradicts the proportional, tiered design. (d) describes only the systemic-risk rule for general-purpose models, a separate track, not the whole Act.

Q2 — (d). No consent was asked (a), there is no contract with the people whose pages were scraped (b), and nobody's life is at stake (c). Controllers of web-scraped training data typically claim legitimate interests, which is only valid together with a documented balancing test, safeguards (such as filtering of personal data), and respect for objections.

Q3 — (b). Article 3 of the Digital Single Market Copyright Directive covers research organizations with no opt-out; Article 4 covers everyone else but yields to a machine-readable rights reservation. (a) ignores both lawful access and the opt-out; (c) is nonsense; (d) is false — temporary copies for mining are exactly what the exception permits.

Q4 — (b). Automation bias is the human over-trusting the automated suggestion and substituting it for their own judgment. (a) is statistical bias from data, (c) is misuse, (d) is a metric side effect (objective misspecification).

Q5. A high-risk artificial-intelligence system is one that is a safety component of a regulated product or falls into a listed sensitive area (biometrics, critical infrastructure, education, employment, essential private and public services such as credit scoring, law enforcement, migration, justice), and is therefore subject to strict obligations before and after market entry. Any three of: risk-management system; data governance with bias examination; technical documentation and logging; transparency and instructions to deployers; designed-in human oversight; accuracy, robustness and cybersecurity; conformity assessment with registration and post-market monitoring.

Q6. Data minimization is the principle that only personal data that is necessary for the specified purpose may be collected and retained. Web-scale pretraining inverts this logic: it collects as much data as possible for a purpose that is defined only vaguely ("train a general model"), so necessity for a specific purpose is hard to demonstrate; mitigations are filtering of personal data, deduplication, opt-out handling, and documented data governance.

Level 2 Q7. Cause: a database stores data as addressable records, but training compresses the corpus into shared parameters; no weight "belongs" to one person. Mechanism: gradient updates from millions

of examples are superimposed in the same weights, so one person's influence is distributed across the whole network and cannot be located or subtracted exactly; rare repeated strings may additionally be memorized verbatim. Consequence: exact erasure requires retraining from scratch without the data, which is prohibitively expensive; approximate machine unlearning is hard to verify; output filtering only suppresses regurgitation without removing the learned information — so the right to erasure collides with the technical structure of trained models.

Q8. Cause: a model's behavior is determined by its training data, objective function, evaluation metric, and deployment context — all chosen at design time. Mechanism: if historical data encodes discrimination, faithful optimization reproduces it; if the objective is a proxy (engagement, similarity to past hires), the model optimizes the proxy including its harmful correlations; accuracy metrics do not measure fairness, so the problem stays invisible. Consequence: no post-release patch fixes this, because nothing is "broken" in the code — the system does exactly what it was designed to do; the remedy is changing the design: data curation, objective and metric selection, per-group evaluation, and deployment constraints.

Q9. Cause: humans over-trust automated outputs (automation bias), especially under time pressure and high system accuracy. Mechanism: verifying is cognitively expensive while agreeing is cheap (cognitive offloading); the suggestion anchors the reviewer; responsibility feels diffused ("the system decided"); and when errors are rare, vigilance decays, so reviewing degenerates into rubber-stamping. Consequence: the human approves precisely the rare wrong outputs that oversight was meant to catch, providing legal cover without safety benefit; effective oversight needs design support — judgment recorded before seeing the suggestion, calibrated confidence display, catch trials, workload limits, and dual control for irreversible actions.

Level 3 Q10 — model answer. Classification (technique): Step 1: RecruitRank is an artificial-intelligence system in scope (machine-based, infers rankings from application data). Step 2: it matches no prohibited practice (no social scoring across contexts, no manipulation, no biometric identification). Step 3: it operates in the employment area — recruitment and selection of natural persons — which is a listed high-risk use, and ranking applicants is not a narrow preparatory task because it materially shapes who is invited. Therefore RecruitRank is a high-risk artificial-intelligence system. Steps 4–5 are moot; step 7 still applies (applicant data is personal data, so the General Data Protection Regulation applies in parallel, including the right not to be subject to a solely automated rejection). Roles and obligations (justification): the vendor is the provider: risk-management system, data governance with bias examination of training data, technical documentation, logging capability, instructions for use, designed-in human oversight, accuracy and robustness testing, conformity assessment and registration. Each customer company is a deployer: use according to instructions, competent human oversight of the ranking (recruiters must be able to question and override it), relevant input data, log keeping, informing applicants about the use of the system. If a customer substantially modifies the system (for example retrains it on its own outcomes and rebrands it), it becomes a provider itself. Trade-off: full compliance (documentation, bias audits, human review of rankings) raises cost and slows hiring throughput — exactly the efficiency the product was bought for; skipping it, however, exposes both parties to fines and discrimination liability.

Q11 — model answer. Computation (technique): selection rate of group M = 88 / 400 = 0.22; selection rate of group N = 45 / 125 = 0.36. Disparate impact ratio = 0.22 / 0.36 = 0.61. Verdict: $0.61 < 0.80 \rightarrow$ the model fails the eighty percent rule; adverse impact against group M (equivalently, 0.22 is below the threshold $0.80 \times 0.36 = 0.29$). Demographic parity difference = $0.36 - 0.22 = 0.14$. Mitigation (justification): first audit why the gap exists (feature correlations with group membership, unrepresentative training data); a standard technique is reweighing or resampling the training data,

or applying group-aware threshold adjustment so that selection rates equalize; also check equalized odds to see whether qualified group-M applicants are being missed (true-positive-rate gap). Trade-off: enforcing demographic parity typically costs some predictive accuracy and can conflict with calibration when base rates differ between the groups; threshold adjustment is also legally sensitive because it treats groups explicitly differently — so document the fairness goal and the justification.

Level 4 Q12 — model answer. 1. Full retraining without the data — the only exact remedy: remove the person's records from the corpus and retrain. Limits: months of time and enormous compute cost per request; infeasible as a per-person process; at best batched into periodic retraining cycles with a "removal list". 2. Machine unlearning — approximate methods (influence-based parameter updates, fine-tuning to forget, sharded training such as the slice-and-aggregate approach where only affected shards are retrained). Limits: guarantees are weak for large language models; verifying that the influence is gone is an open research problem; aggressive unlearning degrades unrelated capabilities; sharded designs must exist before training. 3. Output filtering / suppression — block generations containing the person's data, add refusal behavior for queries about them. Limits: the information remains in the weights (the model still "knows"); filters are circumventable by paraphrase and prompt attacks; this mitigates the manifestation, not the processing. Defensible strategy: prevent at ingestion (filtering of personal data, deduplication, opt-out honoring, provenance records so affected data can be located); on receiving a request: remove the data from all retrievable stores (corpus, retrieval indexes, logs, embeddings), add output filters immediately, schedule the removal for the next retraining or unlearning cycle, and document the process — combining immediate symptom control with eventual removal is the currently defensible reading of the right to erasure.

Q13 — model answer. From the definitions, with base rate p, the false positive rate satisfies the identity FPR = [p / (1 − p)] × [(1 − PPV) / PPV] × TPR, because positive predictive value = true positives / (true positives + false positives), true positives = p × TPR × n and false positives = (1 − p) × FPR × n. Group A: FPR(A) = (0.50 / 0.50) × (0.20 / 0.80) × 0.75 = 1.00 × 0.25 × 0.75 = 0.19 (0.1875). Group B: FPR(B) = (0.20 / 0.80) × (0.20 / 0.80) × 0.75 = 0.25 × 0.25 × 0.75 = 0.05 (0.0469). The false-positive-rate gap is 0.19 − 0.05 = 0.14 ≠ 0, so equal calibration plus equal true positive rate forces unequal false positive rates whenever base rates differ — equalized odds and calibration are jointly unsatisfiable for an imperfect classifier (Chouldechova's impossibility result). Implication: fairness-metric choice is a normative decision: for lending one may prioritize calibration (scores mean the same thing in both groups), for criminal-justice risk scores one may prioritize equal false positive rates (equal burden of false accusations); engineers must state and justify the chosen metric rather than claim "the model is fair".

Level 5 Q14 — solution (verified run).

```
def fairness_report(records, threshold=0.80):
    """Per-group selection rate, TPR, FPR; disparate-impact ratio;
    demographic-parity difference; equalized-odds gaps with verdicts."""
    groups = sorted({r["group"] for r in records})
    stats = {}
    for g in groups:
        rows = [r for r in records if r["group"] == g]
        positives = [r for r in rows if r["y_true"] == 1]
        negatives = [r for r in rows if r["y_true"] == 0]
        sel = sum(r["y_pred"] for r in rows) / len(rows)
        tpr = (sum(r["y_pred"] for r in positives) / len(positives)) if positives else 0.0
        fpr = (sum(r["y_pred"] for r in negatives) / len(negatives)) if negatives else 0.0
```

```
        stats[g] = {"selection_rate": sel, "tpr": tpr, "fpr": fpr, "n": len(rows)}

    rates = [stats[g]["selection_rate"] for g in groups]
    di = min(rates) / max(rates) if max(rates) > 0 else 0.0
    dpd = max(rates) - min(rates)
    tprs = [stats[g]["tpr"] for g in groups]
    fprs = [stats[g]["fpr"] for g in groups]
    tpr_gap = max(tprs) - min(tprs)
    fpr_gap = max(fprs) - min(fprs)

    for g in groups:
        s = stats[g]
        print(f"Group {g}: n={s['n']}, selection rate={s['selection_rate']:.2f}, "
              f"TPR={s['tpr']:.2f}, FPR={s['fpr']:.2f}")
    print(f"Disparate-impact ratio        = {di:.2f} -> "
          f"{'FAILS the 80 percent rule' if di < threshold else 'passes the 80 percent rule'}")
    print(f"Demographic-parity difference = {dpd:.2f}")
    print(f"Equalized-odds gaps: TPR gap = {tpr_gap:.2f}, FPR gap = {fpr_gap:.2f} -> "
          f"{'equalized odds VIOLATED' if max(tpr_gap, fpr_gap) > 0.10 else 'equalized odds approximately
    return di, dpd, tpr_gap, fpr_gap


data = (
    # Group A: 5 actual positives (4 predicted positive), 5 actual negatives (1 predicted positive)
    [{"group": "A", "y_true": 1, "y_pred": 1}] * 4 +
    [{"group": "A", "y_true": 1, "y_pred": 0}] * 1 +
    [{"group": "A", "y_true": 0, "y_pred": 1}] * 1 +
    [{"group": "A", "y_true": 0, "y_pred": 0}] * 4 +
    # Group B: 5 actual positives (3 predicted positive), 5 actual negatives (0 predicted positive)
    [{"group": "B", "y_true": 1, "y_pred": 1}] * 3 +
    [{"group": "B", "y_true": 1, "y_pred": 0}] * 2 +
    [{"group": "B", "y_true": 0, "y_pred": 0}] * 5
)

fairness_report(data)
```

Verified output:

```
Group A: n=10, selection rate=0.50, TPR=0.80, FPR=0.20
Group B: n=10, selection rate=0.30, TPR=0.60, FPR=0.00
Disparate-impact ratio        = 0.60 -> FAILS the 80 percent rule
Demographic-parity difference = 0.20
Equalized-odds gaps: TPR gap = 0.20, FPR gap = 0.20 -> equalized odds VIOLATED
```

Interpretation: group B's selection rate (0.30) is only 60 percent of group A's (0.50), failing the eighty percent rule; the true-positive-rate gap of 0.20 shows that qualified members of group B are missed twice as often, so equalized odds is violated as well.

বাংলা ব্যাখ্যা: কোডের যুক্তি সহজ: প্রতিটি দলের জন্য তিনটে হার বের করো — selection rate (সবার মধ্যে কতজন পজিটিভ পেল), TPR (সত্যিকারের যোগ্যদের মধ্যে কতজন ধরা পড়ল), FPR (অযোগ্যদের মধ্যে কতজন ভুল করে পজিটিভ পেল)। তারপর min/max

অনুপাত দিয়ে disparate impact, বিয়োগ দিয়ে parity difference আর odds gap। পরীক্ষায় কোড লিখতে বললে এই কাঠামোটাই দাও — group-wise লুপ, ভাগ করার আগে শূন্য-চেক, আর শেষে স্পষ্ট verdict প্রিন্ট।

Final Cheat Sheet

One-line definitions to recall under time pressure: - Lawful basis — one of six legal grounds; web scraping → legitimate interests + balancing test. - Right to erasure vs weights — retrain (exact, costly) / unlearn (approximate) / filter outputs (masks only). - Text and data mining exception — research: no opt-out; commercial: machine-readable opt-out wins. - Provider / deployer / user — builds / operates / interacts; substantial modification promotes a deployer to provider. - Risk pyramid — banned · high (full obligations) · limited (transparency) · minimal (nothing). - General-purpose model duties — documentation, copyright policy, training-content summary; systemic risk above 10^25 floating-point operations adds evaluations, red-teaming, incident reporting, cybersecurity. - Eighty percent rule — DI = min selection rate ÷ max selection rate; < 0.80 → adverse impact. - Equalized odds — equal true positive rate and false positive rate per group; report the gaps. - Automation bias — over-trust → rubber-stamping; fix with judgment-first review, confidence display, catch trials, dual control. - Dual use vs misuse — inherent capability duality vs deliberate abuse; red-team, restrict, watermark, stage releases.

Common traps: 1. Confusing the General Data Protection Regulation (regulates personal data) with the Artificial Intelligence Act (regulates systems by use-case risk) — most real systems must satisfy both. 2. Treating pseudonymized data as anonymous — it is still personal data. 3. Believing "the model does not store data" — memorization and regurgitation are real. 4. Treating bias as purely a data problem — objectives, metrics, and deployment context also inject it. 5. Counting any human-in-the-loop as oversight — automation bias makes naive review ineffective. 6. Computing the disparate impact ratio the wrong way around — it is always min ÷ max, so it lies in (0, 1].

End of Chapter 7.