Open Access | https://doi.org/https://doi.org/10.5281/zenodo.17940942

Safe and secure use of AI in research projects

A living open educational resource, workbook, and workshop

Renat Shigapov

Abstract

AI tools increasingly support all stages of research projects. At the same time, their use is constrained by ethical principles, research-integrity standards, and governance (legal and regulatory) requirements. This workshop introduces these constraints, presents the risks of AI use in research, and provides practical strategies for mitigating them across the entire research lifecycle. Both unintentional harms (AI safety) and deliberate threats (AI security) will be addressed. Participants will learn how to use AI tools safely, securely, and in compliance with ethical, integrity, and governance expectations.

Keywords AI safety, AI security, research ethics, research integrity, research governance, risk management, defense in depth

1. Introduction

1.a. Basic definitions

There exist many definitions of AI. None of them are perfect. But when we talk about AI, we often mean AI systems.

AI system is a special type of information technology (IT) system consisting of hardware, software, data, AI model(s) and networks, which can (not necessarily deterministically) generate outputs based on data it receives and data it was trained on.

Local AI system runs fully on a device you control, meaning you also have more control over its safety and security.

Cloud AI system runs on remote infrastructure managed by a provider. Using a cloud AI system means trusting a complex supply chain of hardware, software, data, models, and networks that you do not control.

An Agentic AI system is an AI system that can set or interpret goals, plan actions, and carry out tasks with minimal human intervention or autonomously. To achieve these goals, it may interact with other IT systems via networks or execute code and communicate with software components locally on an IT system. Agentic AI may use multiple models, including a router model that decides which model or tool to use next.

An AI model consists of the model architecture, model parameters (including weights) and inference code for running the model. AI weights are the set of learned parameters that overlay the model architecture to produce an output from a given input. Source: https://opensource.org/ai/open-source-ai-definition

Workbook

Published Dec 16, 2025

Data Availability
GitHub: https://github.com/shigapov/safe_ai/. Zenodo: Shigapov (2025a).

An Open Source AI is one made freely available with all necessary code (for data pre-processing; training, validation, and testing; inference; supporting libraries and tools), data (datasets, data card, technical report, and research paper) and model (architecture and parameters) under legal terms approved by the Open Source Initiative. Source: https://opensource.org/ai/open-source-ai-definition

Generative AI system is a special type of AI system, which was trained to generate content (text, image, audio, video, code, etc.).

(Pre-)training "Developers feed models massive amounts of diverse data – such as text, code, and images – to instil general knowledge. Pre-training produces a 'base model'. This is a highly compute-intensive process." See (Bengio et al., 2025).

Fine-tuning „Developers further train the base model to optimise it for a specific application or make it more useful generally. This is typically done with the help of a large amount of human-generated feedback. This is a moderately compute-intensive and highly labour-intensive process." See (Bengio et al., 2025).

A Retrieval-Augmented Generation (RAG) system is an AI system that retrieves relevant information from either static sources (e.g., a local vector database) or dynamic sources (e.g., internet search), and uses this retrieved information to augment the model's internal knowledge when generating outputs. The model parameters are not updated in RAG.

1.b. Reproducibility and replicability of AI systems

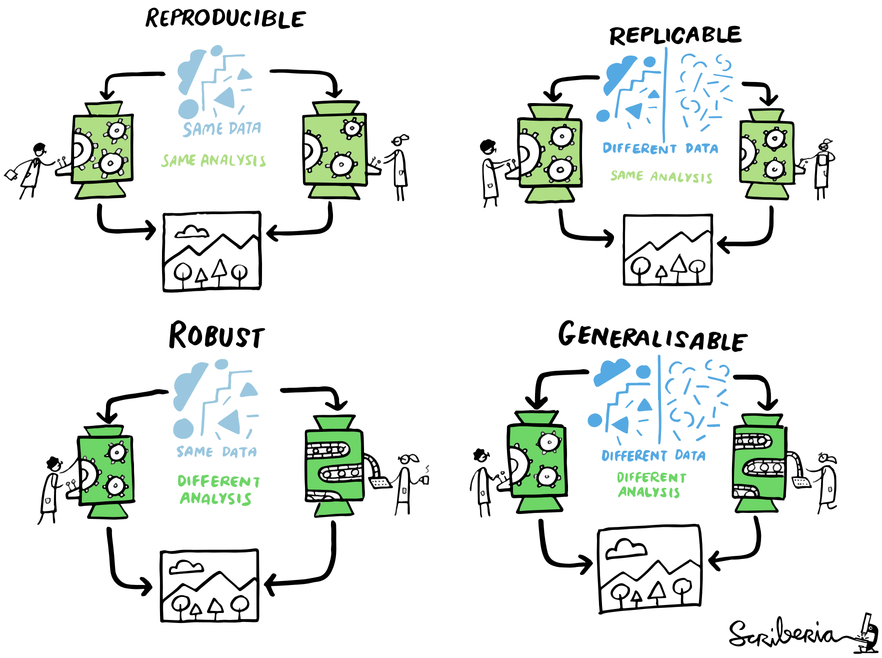


Image source: (Community & Scriberia, 2024).

Various aspects influence reproducibility of outputs in AI systems:

1. Pseudorandomness in computers:

Exercise 1: Testing reproducibility of an AI system at your laptop

- Let’s choose an open-weight model from from the European Open Source AI index (https://osai-index.eu/the-index). Olmo-3-1125-32B is the current number one. Test it via https://huggingface.co/allenai/Olmo-3-1125-32B.
- If that model is too big for your laptop, let’s choose a smaller Olmo 7b model provided by ollama as well. First, install ollama https://ollama.com. Then, run “ollama run olmo2:7b”.
- Adapt the script on the right side and test multiple runs with different parameters. Check the definitions of parameters and find their combination providing reproducible outputs.

    - Computers use pseudorandom sequences, not true randomness.
    - A random seed initializes the sequence.

2. GenAI models choose tokens by sampling from probability distributions:

    - Temperature sampling: adjusts shape of the distribution. Higher temperature flattens distribution, making generator more random. Lower temperature sharpens distribution, making generator more deterministic.
    - Top-K sampling: samples from the K most likely tokens.
    - Top-P sampling: samples from the smallest set of tokens whose cumulative probability ≥ P.

3. Floating-point non-associativity $(a+b)+c \neq a+(b+c)$ in GPUs, TPUs, and CPUs:

    - He (2025)
    - Yuan et al. (2025)

To access your local ollama instance via API, you can use the code:

```python
import requests


def query(prompt: str):
    url = "http://localhost:11434/api/generate"
    payload = {
        "model": "olmo2:7b",
        "prompt": prompt,
        "stream": False,
        "options": {
            "seed": 42,
            "temperature": 1,
            "top_k": 10,
            "top_p": 1,
        }
    }

    resp = requests.post(url, json=payload)
    resp.raise_for_status()
    data = resp.json()
    return data["response"]
```

For your reproducibility tests, you can adapt:

```
from collections import Counter

def test_reproducibility(prompt: str, runs: int = 10):
    outputs = []

    print(f"Running reproducibility test for prompt:\n{prompt}\n")

    for i in range(runs):
        out = query(prompt)
        outputs.append(out)
        print(f"Run {i+1}:")
        print(out.strip()[:300] + ("..." if len(out) > 300 else ""))
        print("-" * 40)

    # Count unique outputs
    counts = Counter(outputs)
    print("\n=== Summary ===")
    print(f"Total runs: {runs}")
    print(f"Unique outputs: {len(counts)}\n")

    for text, count in counts.items():
        print(f"Occurrences: {count}")
        print("Sample:")
        print(text.strip()[:300] + ("..." if len(text) > 300 else ""))
        print("-" * 40)


# Run the reproducibility test
test_reproducibility("What is reproducibility of an LLM?", runs=10)
```

1.c. Safe and secure AI system

There are no zero-risk AI systems. The only fully safe and fully secure AI system is the one that is never used. But modern research and everyday life rely increasingly on AI systems. Therefore, it is important to identify AI risks and learn how to manage them.

Safety and security of AI system include safety and security of its components: hardware, software, data (including training, validation, testing, augmented, input, output, and database), models, and networks. This makes the topic very complex.

Safe and secure use of AI in research (projects) is shaped by ethical, integrity, and governance (legal and regulatory) requirements and expectations. They are complex as well.

1.d. AI Safety and AI Security

According to (Lin et al., 2025) AI safety considers risks from accidental or unintended behaviors. AI Security deals with risks from intentional or adversarial actions by malicious actors

1.e. *Safety and security in context of research*

The trinity of good research: Research ethics, research integrity, and research governance (Kolstoe & Pugh, 2024). Safety and security are parts of all of them.

1.f. *Risk management*

OWASP AI Risk Analysis page describes the Risk Management Framework:

- Identifying Risks
- Evaluating Risks by Estimating Likelihood and Impact
- Deciding What to Do (Risk Treatment):
  - Risk Mitigation,
  - Transfer,
  - Avoidance,
  - Acceptance.
- Risk Communication and Monitoring

We cannot treat risks unless we can identify them. We identify risks within research ethics, research integrity, and research governance. The general risks in AI safety and AI security will be also introduced. In Part 2 we do risk management across the entire research lifecycle.

1.g. *Defense in depth*

Defense in depth is a risk-mitigation approach that relies on multiple, overlapping, and complementary layers of safeguards, so that if one control fails, others remain effective.

Key assumption: any single mitigation measure can fail.

Dimensions of defense in research (projects) for an AI use case:

- **People:** individual researchers, research teams, project partners, and participants
- **Infrastructure/Technology:** hardware, software, data, models, networks, and the entire supply chain
- **Processes:** research process including research ethics, research integrity, and research governance

1.h. *Risk management framework adapted to the trinity of risks in research (projects)*

- **Identifying Risks:** Recognizing potential risks related to research ethics, research integrity, and research governance.
- **Evaluating Risks:** Subjectively as low, medium, high.
- **Risk Treatment:** Choosing an appropriate strategy to address the risk:
  - Risk Mitigation,
  - Transfer (to another party),

  - Avoidance (eliminating the source),
  - Acceptance.
- **Risk Communication and Monitoring:** Regularly sharing risk information with stakeholders. Creating a Risk Register, a list of risks and their attributes (e.g. severity, treatment plan, ownership, status, etc).

This framework is adapted specifically for research context from the Risk management framework: https://owaspai.org/goto/riskanalysis/

> **Note**
>
> Note: Any mitigation measure can fail. Apply defense in depth to all three dimensions (people, technology, and processes) in Risk Mitigation.

1.i. Questions

Before you use any AI system, ask yourself the following questions:

- Is the hardware safe and secure?
- Is the software safe and secure?
- Is the data (including training, validation, testing, augmented, input, and output) safe and secure?
- Is the model safe and secure?
- Are the networks safe and secure?

There are three kinds of AI systems:

- **Cloud AI:** Choose a compliant provider (check certifications and trust portals) + use case
- **Local AI (self-built):** You are responsible for safety, security, and use case
- **Hybrid (Local AI with third-party open-weight models):** Shared responsibility (see licenses)

2. 1. Theory

In the theoretical part we introduce:

- Freedom of research under ethical, integrity, and governance (legal and regulatory) constraints
- AI risks, AI safety, and AI security
- Risk management

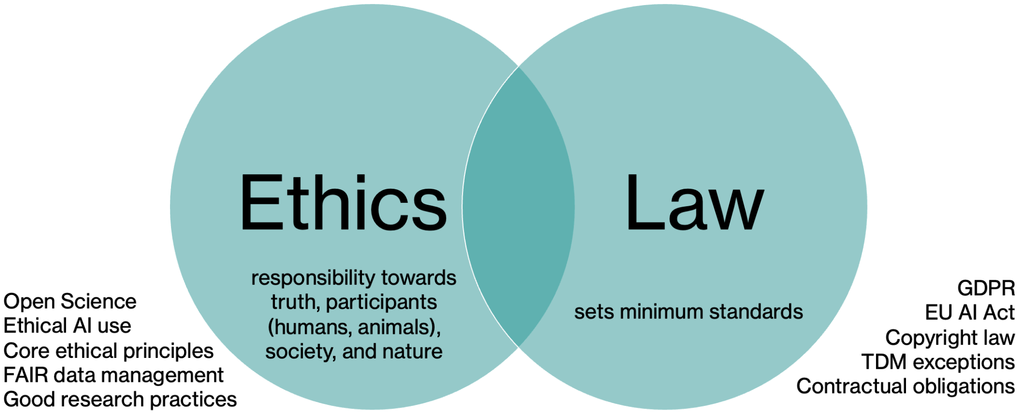
2.a. *Freedom of research under constraints*

2.a.i. *Freedom of scientific research:*

- "The arts and scientific research shall be free of constraint. Academic freedom shall be respected." Art. 13 [Freedom of the arts and sciences], EU Charter of Fundamental Rights, https://fra.europa.eu/en/eu-charter/article/13-freedom-arts-and-sciences
- "Arts and sciences, research and teaching shall be free." Art. 5 [Freedom of expression, arts and sciences], Basic Law for the Federal Republic of Germany. https://www.gesetze-im-internet.de/englisch_gg/englisch_gg.html
- Art. 8, EU Charter of Fundamental Rights, https://fra.europa.eu/en/eu-charter/article/13-freedom-arts-and-sciences:
  - "Protection of personal data
    1. Everyone has the right to the protection of personal data concerning him or her.
    2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.
    3. Compliance with these rules shall be subject to control by an independent authority."
- Art. 27, Universal Declaration of Human Rights, https://www.un.org/en/about-us/universal-declaration-of-human-rights: "Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author."

2.a.ii. *The duality of good research:*

It is not uncommon to introduce safe and secure AI systems through the lenses of ethics and law (Enrico Glerean, 2025).



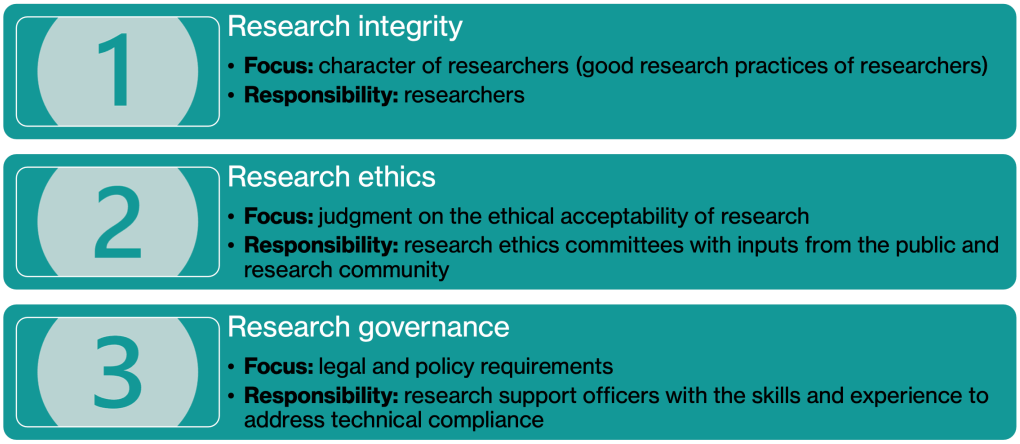
In context of research ethics means responsibility towards truth, participants (humans, animals), society, and nature. It includes Open Science, Ethical AI use, core ethical principles, FAIR data management, and good research practices.

In contrast to ethics, law sets minimum standards. It includes GDPR, EU AI Act, Copyright law, TDM exceptions, contractual obligations, etc.

2.a.iii. *The trinity of good research:*

The trinity of good research was introduced by Kolstoe & Pugh (2024). They gave clear definitions of research ethics, research integrity, and research governance. They differ in their focus and corresponding responsibilities.

While research integrity focuses on responsibility towards truth and researcher's behaviour, research ethics deals with responsibility towards participants, society, and the environment.



1 Research integrity

- **Focus:** character of researchers (good research practices of researchers)
- **Responsibility:** researchers

2 Research ethics

- **Focus:** judgment on the ethical acceptability of research
- **Responsibility:** research ethics committees with inputs from the public and research community

3 Research governance

- **Focus:** legal and policy requirements
- **Responsibility:** research support officers with the skills and experience to address technical compliance

2.a.iv. *Too many values, principles, laws, regulations…:*

Researchers face variety of (partially contradicting) values, principles, codes of conduct, guidelines of good practices, ethics frameworks, laws, regulations, institutional policies, funder requirements, contractual obligations, disciplinary norms, open science mandates, data protection rules, security requirements, journal data and AI policies.

It feels impossible to navigate.

2.a.v. *Pragmatic approach: Remember responsibilities:*

To conduct responsible high-quality research, researchers need supporting infrastructure. Research organizations are responsible for providing infrastructure, legal certainty, ethical oversight, data protection, and secure environments. Researchers are supported by ethics committees, ombudspersons, research data management teams, legal & data protection offices, IT security teams, Open Science offices, export control offices, etc.

Hint

Use these support structures early, not when problems appear.

Research ethics:

Should we do this project? Values, principles, social responsibility, impact on humans, animals, society, and nature. GET ethical approval from research ethics committees for your research project.

Note

Independently of ethical approval, your own ethical self-assessment of your research project can significantly improve its quality

Research integrity:

How should we behave? Responsibility, honesty, rigor, transparency, FAIR & Open Science, following good research practices, avoiding questionable research practices & scientific misconduct. ACT responsibly.

Research governance:

What must we comply with? Laws (GDPR, copyright, EU AI Act), policies, funding rules, contracts, licenses, and agreements. GET help from research support units for your research project.

Note

Expect to hear “on a case-by-base basis” and that’s fine.

2.a.vi. *Pragmatic approach: Remember the goal:*

Research ethics, research integrity, and research governance are not constraints but foundations for excellent research. If we agree on that, compliance becomes a quality tool, not paperwork.

2.b. Research Ethics

2.b.i. *Research ethics: Who defines that?:*

- International Ethical Frameworks define the values and principles of ethics (e.g., UNESCO ethical frameworks)
- Disciplinary Norms and Communities define what is ethically sensitive or ethically problematic in their own context. These norms guide ethical evaluation (e.g., the Declaration of Helsinki by the World Medical Association)
- Ethics Committees or Institutional Review Boards check what is ethically permissible in concrete research projects
- Funders may require ethical (self-)assessment and review (e.g., EU grants from three EU programs such as Horizon Europe, Digital Europe, and European Defence Fund)

2.b.ii. *Main research ethics documents for using AI in research projects:*

- UNESCO Recommendation on the Ethics of Artificial Intelligence (https://unesdoc.unesco.org/ark:/48223/pf0000381137)
- OECD Principles on Artificial Intelligence (https://oecd.ai/en/ai-principles)
- EU High-Level Expert Group on AI – Ethics Guidelines for Trustworthy AI (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai)
- EU Grants: How to complete your ethics self-assessment https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf

2.b.iii. *Research ethics in Germany:*

A collection of best practices in research ethics for reseachers and research committees can found at the website of the German Data Forum (RatSWD): https://www.konsortswd.de/en/topics/best-practices-research-ethics

2.b.iv. *UNESCO recommendation on the ethics of AI:*

- A human rights approach to AI
- 4 core values
    - Respect, protection and promotion of human rights and fundamental freedoms and human dignity
    - Environment and ecosystem flourishing
    - Ensuring diversity and inclusiveness
    - Living in peaceful, just and interconnected societies
- 10 principles

Source: (United Nations Educational, Scientific and Cultural Organization (UNESCO), 2022)

UNESCO recommendation: 10 principles:

1. Proportionality and Do No Harm (risk assessment, choosing appropriate AI tools)
2. Safety and security (risks should be addressed, prevented and eliminated)
3. Fairness and non-discrimination
4. Sustainability
5. Right to Privacy, and Data Protection (legal, ethical, and technical compliance)
6. Human oversight and determination (responsibility lies on people or legal entities)
7. Transparency and explainability (a need to balance with privacy, safety, and security)
8. Responsibility and accountability (AI actors and Member States)
9. Awareness and literacy (based on impact on human rights and access to rights, on the environment and ecosystem)
10. Multi-stakeholder and adaptive governance and collaboration

2.b.v. *The Ethics of Artificial Intelligence by Floridi:*

Floridi (2023) considers five ethical principles of AI:

1. **Beneficence (“do only good”):** Promoting Well-Being, Preserving Dignity, and Sustaining the Planet
2. **Nonmaleficence (“do no harm”):** Privacy, Security, and ‘Capability Caution’
3. **Autonomy:** The Power to ‘Decide to Decide’
4. **Justice:** Promoting Prosperity, Preserving Solidarity, Avoiding Unfairness
5. **Explicability:** Enabling the Other Principles through Intelligibility and Accountability

2.b.vi. *Ethics guidelines for trustworthy AI:*

- Trustworthy AI should be lawful, ethical, and robust.
- Trustworthy AI systems should meet 7 requirements:

    1. **Human agency and oversight:** AI systems should empower human beings
    2. **Technical Robustness and safety:** AI systems need to be resilient and secure
    3. **Privacy and data governance:** Privacy and data protection, adequate data governance
    4. **Transparency:** The data, system and AI business models should be transparent
    5. **Diversity, non-discrimination and fairness:** Unfair bias must be avoided
    6. **Societal and environmental well-being:** AI systems should benefit all human beings and environment
    7. **Accountability:** Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes

2.b.vii. *Tool: ALTAI (The Assessment List for Trustworthy Artificial Intelligence):*

ALTAI is the Assessment List for Trustworthy AI.

Exercise 1: ALTAI

- Register at https://altai.insight-centre.org/Assessment and make an assessment of an AI system.
- Alternatively, download the PDF-file with the assessment list https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342 and assess your AI system.

Exercise 2: Question and discussion

What experiences have you had with Ethics Committees or Institutional Review Boards regarding the use of AI in research?

2.b.viii. *Ethical self-assessment in EU grants:*

See https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf

"Any use of AI systems or techniques should be clearly described in the project and you must demonstrate their technical robustness and safety (they must be dependable and resilient to changes)."

2.b.ix. *Tool: Data protection decision tree:*

Exercise 3: Data protection decision tree

- Apply the data protection decision tree https://ec.europa.eu/assets/rtd/ethics-data-protection-decision-tree/index.html to your use case with an AI system.

Exercise 4: Question and discussion

What experiences have you had with ethical (self-)assessment for European grants regarding the use of AI in research?

2.c. Research Integrity

2.c.i. *Research integrity: Who defines that?:*

- International, national, funders, and university Codes of Conduct for Research Integrity (e.g., ALLEA European Code of Conduct, Singapore Statement on Research Integrity, DFG's "Guidelines for Safeguarding Good Research Practice")
- Research communities define domain-specific integrity standards
- Local integrity offices, local ombudspersons, and Ombuds Committee for Research Integrity in Germany are points of contact for researchers to report and seek advice on issues related to research integrity
- Publishers and journals create policies related to research integrity (data policy, AI policy, general guidelines for authors, etc.)
- Committee on Publication Ethics issues guidelines for authors and reviewers

2.c.ii. *Main research integrity documents for using AI in research:*

General documents which do not specifically mention „AI":

- ALLEA European Code of Conduct for Research Integrity (https://allea.org/code-of-conduct)
- Singapore Statement on Research Integrity (https://www.wcrif.org/statement)
- DFG. (2025). Guidelines for Safeguarding Good Research Practice. Code of Conduct. Deutsche Forschungsgemeinschaft (2025) AI-specific documents:
- Living guidelines on the responsible use of generative AI in research for researchers, research organizations, and research funding organizations (https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en)
- AI policies and recommendations of research organizations (e.g., Helmholtz: https://www.helmholtz.de/assets/helmholtz_gemeinschaft/Downloads/Helmholtz_Recommendations_on_use_of_AI_Version_1.0.pdf)
- AI policies of funders, journals, and publishers

2.c.iii. *Research misconduct in European Code of Conduct:*

According to (ALLEA, 2023):

Research misconduct is traditionally defined as fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results:

- Fabrication is making up data or results and recording them as if they were real.
- Falsification is manipulating research materials, equipment, images, or processes, or changing, omitting, or suppressing data or results without justification.
- Plagiarism is using other people's work or ideas without giving proper credit to the original source.

2.c.iv. *DFG and research misconduct: Scope and Definitions:*

The following scope and definitions can be found in DFG, "Rules of Procedure for Dealing with Scientific Misconduct", https://www.dfg.de/resource/blob/339200/dfg-80-01-v0524-en.pdf

"§1 (2): These Rules of Procedure apply if the respondent is one of the following with regard to the allegation:

1. A grant applicant to the DFG
2. A grant recipient funded by the DFG,
3. Individuals with a high level of scientific responsibility in connection with funding proposals submitted by higher education institutions or non-university research institutions,
4. individuals reviewing a proposal for the DFG or
5. A member of a DFG committee or a committee supported by the DFG in administering funding instruments who participates in advisory, review, evaluation or decision-making procedures."

"§2: Scientific Misconduct (1) ^{1}An individual pursuant to § 1 (2) nos. 1-3 commits scientific misconduct if they do any of the following in particular, either intentionally or with gross negligence:

1. make misrepresentations (§ 3),
2. appropriate others' research achievements without justification (§ 4),
3. interfere with others' research (§ 5),
4. participate in the scientific misconduct of others by way of co-authorship (§ 6) or
5. neglect their supervisory duties (§ 7). ^{2}Anyone who intentionally participates in the misconduct of others is also guilty of scientific misconduct (§ 8). (2) A person pursuant to § 1 (2) nos. 4 and 5 commits scientific misconduct if they do any of the following, either intentionally or with gross negligence:
6. breach confidentiality (§ 9),
7. fail to disclose circumstances that give rise to the appearance of conflict of interest (§ 10) or
8. inadmissibly give unfair preferential treatment to others (§ 11). "

2.c.v. *Spectrum of questionable research practices:*

Although usually questionable research practices are defined as something less dangerous as scientific misconduct, Kolstoe (2023) proposed to define a spectrum of questionable practices:

- QRP is "a spectrum of behaviours, ranging from honest errors and mistakes at one end, through to more serious behaviours at the other".
- "Everyone involved in research may at times engage in QRPs, and thus it is up to everyone involved in research to recognise and address the problem in their own, as well as others' research".

QRPs with AI don't just repeat traditional integrity risks — they amplify them. Shigapov (2025b)

A lack of AI literacy may unintentionally lead to the whole spectrum of questionable research practices. Not necessarily your own AI literacy, but that of a co-author, project partner, or student assistant.

2.c.vi. *Tool: Retraction Watch Database, Blog, etc.:*

Exercise 1: Retraction Watch Database and Blog

- Find the posts at https://retractionwatch.com related to retractions of papers due to use of AI
- Download the retraction watch database https://gitlab.com/crossref/retraction-watch-data and find retracted papers with "AI" in the titles
- Read the list of papers with evidence of ChatGPT-writing https://retractionwatch.com/papers-and-peer-reviews-with-evidence-of-chatgpt-writing. Analyze and discuss the reasons for retraction.

2.c.vii. *Tool: Academ-AI:*

Exercise 2: Academ-AI

- Read the samples of papers with suspected undeclared AI usage in the academic literature at https://www.academ-ai.info. Discuss strong and weak markers of using AI in research.
- Read preprint Glynn (2024) and discuss it with colleagues

2.c.viii. *Tool: Research Integrity Risk Index:*

Exercise 3: Research Integrity Risk Index

- Read preprint Meho (2025) and discuss it with colleagues
- Analyze https://sites.aub.edu.lb/lmeho/ri2. How can it help you to select potential project partners?

2.c.ix. *Living guidelines on the responsible use of generative AI in research for researchers:*

1. Remain ultimately responsible for scientific output.
2. Use generative AI transparently.
3. Pay particular attention to issues related to privacy, confidentiality and intellectual property rights when sharing sensitive or protected information with AI tools.
4. Respect applicable national, EU and international legislation.
5. Continuously learn how to use generative AI tools properly to maximise their benefits, including by undertaking training.

6. Refrain from using generative AI tools substantially in sensitive activities that could impact other researchers or organisations (for example peer review, evaluation of research proposals, etc).

Source: https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en

2.c.x. *Living guidelines on the responsible use of generative AI in research for research organizations:*

The guidelines for research organizations include: "3. Reference or integrate these generative AI guidelines into their general research guidelines for good research practices and ethics. Using these guidelines as a basis for discussion, research organisations openly consult their research staff and stakeholders on the use of generative AI and related policies. Research organisations apply these guidelines whenever possible. If needed, they could be complemented with specific additional recommendations and/or exceptions that should be published for transparency."

2.c.xi. *German FAQ on AI and research integrity:*

- DE: Frisch, K. (2025). FAQ Künstliche Intelligenz und gute wissenschaftliche Praxis - Version 2. Zenodo. Frisch (2025)
- EN: Frisch, Katrin (2025). FAQ Artificial Intelligence and Research Integrity. Version 2. Zenodo. Frisch (2025)
- More resources by Dr. Katrin Frisch on research data and AI in context of research integrity are available at https://ombudsgremium.de/9806/research-data-and-ai/?lang=en

2.c.xii. *Tool: Artificial Intelligence Disclosure (AID):*

Exercise 4: Artificial Intelligence Disclosure (AID)

- Test https://aidframework.org

2.c.xiii. *Tool: AI Attribution Toolkit:*

Exercise 5: AI Attribution Toolkit

- Test https://aiattribution.github.io/interpret-attribution

2.c.xiv. *Tool: GAIDeT Declaration Generator:*

Note

- Test https://panbibliotekar.github.io/gaidet-declaration/index.htm
- Read paper Suchikova et al. (2025) and discuss it with colleagues

2.c.xv. *Tool: Decision tree for responsible application of AI:*

Note

- Apply the decision tree for responsible application of AI to your AI use case: American Association for the Advancement of Science (AAAS). Decision Tree for the Responsible Application of Artificial Intelligence (v1.0): [Online]. (2023). https://www.aaas.org/sites/default/files/2023-08/AAAS Decision Tree.pdf

Exercise 8: Quiz

Who do you believe is responsible for research integrity when AI systems are used?

- Users of AI systems
- Supervisors, PIs, and project managers
- Co-authors
- Project partners
- Institutions and support teams
- Developers and providers of AI systems
- All of the above

2.d. Research Governance

2.d.i. *Research governance: Who defines that?:*

- Lawmakers
  - National: German Bundestag and Bundesrat (e.g., BDSG, Urheberrechtsgesetz)
  - European: European Parliament and Council (e.g, GDPR, EU AI Act, CDSM directive)
- Regulator and government agencies (e.g., European Medicines Agency, European Chemicals Agency, and Data Protection Authorities)
- Courts interpret how the law applies
- International Treaties and Conventions (e.g., UNESCO conventions)
- Institutional Governance Structures (policies, support structures, and processes)
- Contracts (grant agreements, consortium agreements, NDAs, licenses, DSAs, etc.)
- Mechanisms to ensure compliance (procedures and checklists)

2.d.ii. *Main research governance documents for using AI in research projects:*

- Data protection and privacy: GDPR (General Data Protection Regulation), BDSG (Bundesdatenschutzgesetz); see also Talus, Anu. Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models. The European Data Protection Board, 2024. https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf]
- Copyright law: EU copyright law consists of 13 directives and 2 regulations including InfoSoc Directive, CDSM or just DSM (Copyright in the Single Market) directive, Software directive, Database directive, etc.; see also European Union Intellectual Property Office, The development of generative artificial intelligence from a copyright perspective, European Union Intellectual Property Office, 2025, European Union Intellectual Property Office. (2025)
- AI-specific: EU AI Act https://artificialintelligenceact.eu
- Dual use and security-relevant research: Regulation (EU) 2021/821 (EU Dual-Use Regulation), “Manual Export Control and Academia” by BAFA and “Recommendations for Handling Security-Relevant Research” by DFG and Leopoldina
- Contract law: software, data, and model licenses, consortium agreements, grant agreements, industry collaborations, cloud service terms, NDAs, data processing agreements (DPAs), etc.

2.d.iii. *GDPR:*

All citations in this part are from http://data.europa.eu/eli/reg/2016/679/oj, until otherwise stated.

Definitions:

- Art. 4 (1): "'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly".
- Art. 4 (7): "'controller' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data"
- Art. 4 (8): "'processor' means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller"

Non-applicability:

- Recital 26: "Not Applicable to Anonymous Data"
- "The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes."

Principles:

Art. 5 (1) introduces the principles for personal data and its processing:

- Lawfulness, fairness and transparency (process legally, treat people fairly, and explain clearly)
- Purpose limitation (only for specific purposes)
- Data minimization (use only the minimum amount of personal data necessary)
- Accuracy (ensure personal data is correct and keep it updated)
- Storage limitation (do not keep personal data longer than necessary)
- Integrity and confidentiality (ensure appropriate security of the data to prevent loss, misuse, or unauthorised access using appropriate technical or organisational measures) Art. 5 (2) introduces the principle of responsibility of the controller for paragraph 5.
- Accountability

Legal basis or lawfulness:

Art. 6 (1) "Processing shall be lawful only if":

1. Consent
2. Contract
3. Legal obligation
4. Protection of vital interests of a natural person

5. Public task done in public interests
6. Legitimate interest

Note

Legitimate interest is an argument of OpenAI with respect to using scraped, publicly available, personal data for ChatGPT training https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed. See also https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf.

The rights of individuals:

- Right to be informed: You can obtain information about the processing of your personal data.
- Right of access: You can obtain access to the personal data held about you.
- Right to rectification: You can ask for incorrect, inaccurate or incomplete personal data to be corrected.
- Right to erasure: You can request that personal data be erased when it's no longer needed or if processing it is unlawful.
- Right to restriction of processing: You can request the restriction of the processing of your personal data in specific cases.
- Right to data portability: You can receive your personal data in a machine-readable format and send it to another controller.
- Right to object: You can object to the processing of your personal data for marketing purposes or on grounds relating to your particular situation.
- Rights in relation to automated decision-making and profiling: You can request that decisions based on your personal data and that significantly affect you are made by natural persons, not only by computers.

Processing of special categories of personal data:

Art. 9 (1): "Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited."

Art. 9 (2): A selected list of exceptions relevant in research context

- Consent
- Protection of vital interests of a natural person
- Personal data which are manifestly made public by the data subject
- Public interests
- For archiving purposes in the public interest, scientific or historical research purposes or statistical purposes

Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes:

Art. 89 (1):

- Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards […] for the rights and freedoms of the data subject.
- Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation.
- Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner.

Security of processing:

Art. 32 (1) "the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk":

- the pseudonymisation and encryption of personal data;
- the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services;
- the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident;
- a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing.

More important excerpts for research:

Art. 44-50: Personal data may only be transferred outside of the European Economic Area in compliance with the conditions for such transfers laid down in Chapter 5 of the GDPR. The main types of transfer tools include standard data protection clauses (SCCs), binding corporate rules (BCRs), codes of conduct, certification mechanisms, and ad hoc contractual clauses.

Recital 156: "The processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be subject to appropriate safeguards for the rights and freedoms of the data subject"

Data protection impact assessment (DPIA):

- Art. 35 (1): "Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data."

- The purpose of the DPIA is to identify, assess and mitigate any risks to the rights and freedoms of affected (natural) persons that may result from data processing ("The Data Protection Impact Assessment according to Article 35 GDPR. A Practitioner's Manual.", https://publica-rest.fraunhofer.de/server/api/core/bitstreams/e6b91341-71f4-409b-8446-03432231a0d0/content)

Exercise 1: ❓ Question and discussion

Was a data protection impact assessment ever conducted for your research project if you processed personal data using AI?

What are risks as defined in the GDPR?:

- Recital 75: "The risk to the rights and freedoms of natural persons, of varying likelihood and severity, may result from personal data processing which could lead to physical, material or non-material damage"
- „Non-material damages may be of a social, personal, and legal nature" (see „The Data Protection Impact Assessment according to Article 35 GDPR. A Practitioner's Manual"):
    - Social disadvantages
    - Damage to privacy
    - Chilling effects (e.g., a state in which persons refrain from exercising their rights)
    - (Unjustified) interference with rights

GDPR and AI:

GDPR applies to all stages of the AI lifecycle if personal data is processed (including data collection, filtering, and processing; model training, fine-tuning, augmentation, validation, and inference; inputs and outputs of an AI system; data & model archiving)

EDPB opinion on AI models:

- The European Data Protection Board (EDPB) issued opinion on AI models: „GDPR principles support responsible AI" (https://www.edpb.europa.eu/news/news/2024/edpb-opinion-ai-models-gdpr-principles-support-responsible-ai_en)
- "The EDPB considers that, for an AI model to be considered anonymous, using reasonable means, both (i) the likelihood of direct (including probabilistic) extraction of personal data regarding individuals whose personal data were used to train the model; as well as (ii) the likelihood of obtaining, intentionally or not, such personal data from queries, should be insignificant for any data subject. By default, supervisory authorities should consider that AI models are likely to require a thorough evaluation of the likelihood of identification to reach a conclusion on their possible anonymous nature. This likelihood should be assessed taking into account 'all the means reasonably likely to be used' by the controller or another person, and should also consider unintended (re)use or

disclosure of the model" (https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf)
- "When an AI model was developed with unlawfully processed personal data, this could have an impact on the lawfulness of its deployment, unless the model has been duly anonymised." (https://www.edpb.europa.eu/news/news/2024/edpb-opinion-ai-models-gdpr-principles-support-responsible-ai_en)

Note

Exercise 2: Testing memorization of personal data with ChatGPT: Ask ChatGPT about yourself.

Note

Exercise 3: Testing memorization of personal data with open-weight models. To choose an AI model, you could use the European Open Source AI Index. For example, you can start with https://osai-index.eu/model/OLMo.

Tools: Data anonymization at your laptop:

- Existing data anonymization tools often use named entity recognition pipelines in them
  - https://microsoft.github.io/presidio
  - https://github.com/microsoft/presidio
- Alternatively, one could use locally-hosted open-weight model for anonymization. See example.

Data controls in ChatGPT:

- https://chatgpt.com/#settings/DataControls
- https://privacy.openai.com/policies?modal=take-control

Data residency and inference residency in ChatGPT:

- Data residency (storage at rest) controls where customer content is stored when it is saved by the service (for example, chat history, files, and GPT configurations).
- Inference residency (model execution) controls where model inference on customer content runs on GPUs (for example, generating responses, embedding documents), for supported regions.
- Data residency for ChatGPT is currently available in the following regions: Australia, Canada, Europe (EEA + Switzerland), India, Japan, Singapore, South Korea, United Arab Emirates, United Kingdom, and United States
- Who can use it? Eligible API customers and new ChatGPT Enterprise/Edu customers

02.12.2025 (https://help.openai.com/en/articles/9903489-data-residency-and-inference-residency-for-chatgpt): „Inference residency for ChatGPT is currently available for the United States. It requires data residency in the U.S.“ „We plan to expand supported inference residency regions over time and will update this article as new regions become available.”

ChatGPT and GDPR-Compliance:

The GDPR does not mandate data localization, but it outlines strict rules and requirements for processing data outside of the EEA, including adequacy decisions, standard contractual clauses, certifications, and binding corporate rules.

See https://trust.openai.com/?itemUid=45220873-6e51-4dbb-b1b1-37d66ee9ef95

Exercise 4: ❓ Quiz

Who is responsible for GDPR compliance in a research project, if researchers use AI systems with personal data?

- The AI system provider
- The AI system developer
- The researcher alone
- The research institution as controller

Exercise 5: ❓ Quiz

Which AI output can be personal data under GDPR?

- A generated image
- A prediction about an individual
- A summary mentioning identifiable persons
- All of the above

Exercise 6: ❓ Quiz

When is a Data Protection Impact Assessment (DPIA) required?

- Whenever processing is likely to result in a high risk to the rights and freedoms of individuals
- A systematic and extensive evaluation of the personal aspects of an individual, including profiling
- Processing of sensitive data on a large scale
- All of the above

2.d.iv. *Intellectual property rights and copyright:*

Intellectual property rights (IPR) are legal rights that protect creations of the mind such as inventions, publications, datasets, software, designs, and trademarks. Typical IPRs are copyright (publications, presentations, datasets [exception: facts are public domain], and software), patent rights, design rights, trademarks (names and logo), database rights, and trade secrets. In EU countries, copyright protects your intellectual property until 70 years after your death or 70 years after the death of the last surviving author in the case of a work of joint authorship. Outside of the EU, in any country which signed the Berne Convention, the duration of copyright protection can vary but it lasts until at least 50 years after the author's death. See also https://eur-lex.europa.eu/EN/legal-content/summary/copyright-and-related-rights-in-the-information-society.html

Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society (InfoSoc):

- Art. 2-4: Authors and neighboring rightsholders have the reproduction right, the right of communication to the public and the distribution right (see Directive 2001/29/EC InfoSoc).
- Art. 5 (1): There is a mandatory exception to the right of reproduction for certain temporary acts of reproduction which are an integral and essential part of a technological process (temporary copies), and which aim to enable a lawful use or a transmission in a network between third parties by an intermediary, of a work or other subject matter.

Directive (EU) 2019/790 on copyright in the Digital Single Market:

The directive makes it easier to use copyright-protected material for different purposes, mostly related to access to knowledge, by introducing mandatory exceptions to copyright to foster:

- text- and data-mining (TDM);
- digital uses of works for the purpose of illustration for teaching; and
- the preservation of cultural heritage.

TDM is "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations", Art. 2 (2))

Text and data mining for the purposes of scientific research:

Art. 3 of the CDSM:

- (1) Exception for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.

- (2) Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results.
- (3) Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.

Exception or limitation for text and data mining:

Art. 4 of the CDSM:

- (1) Exception for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.
- (2) Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining.
- (3) The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.

Opt-out measures in practice:

No single opt-out mechanism has emerged as the sole standard used by rights holders:

1. Technical measures:
    - The Robots Exclusion Protocol (REP) (Koster et al. (2022))
    - TDM reservation protocol (https://github.com/w3c/tdm-reservation-protocol)
2. Legally-driven measures:
    - unilateral declarations by copyright holders
    - licensing constraints
    - website terms and conditions

Tool: TDM exception decision tree for researchers:

Iacino et al. (2023)

EUIPO's copyright perspective on AI:

Summary of opinion is available at https://www.euipo.europa.eu/en/publications/genai-from-a-copyright-perspective-2025:

- No 'one-size-fits all' solution for copyright holders to protect their rights has emerged yet.

- Instead, different approaches and solutions are developing for copyright holders to protect their rights, and for AI developers to respect their regulatory obligations.
- On the one side, the rights reservation mechanisms for the INPUT phase (related to training AI models), whereby rightsholders can express their opt out from the ‘text and data mining’ (TDM)-exception.
- On the other side, transparency measures exist for the OUTPUT phase that allow the indication and recognition of AI generated content.

The opinion is published in European Union Intellectual Property Office. (2025).

Case Study: LAION v. Kneschke (Hamburg district court, 27 Sept 2024, 310 O 227/23):

- Background:
    - Photographer Robert Kneschke sued LAION (https://laion.ai), a non-profit that creates AI training datasets using web-scraped publicly available images or the Common Crawl.
    - His allegation: LAION reproduced his image without permission while building its dataset https://laion.ai/blog/laion-5b .
- Court’s Decision: LAION’s activity (dataset creation) is protected under Section 60d UrhG (TDM for scientific research) implementing Article 3 CDSM Directive.
- Note: The shared dataset contains only metadata (e.g., URLs), not the images.
- Preliminary conclusion: TDM exception seems to be applicable for dataset creation. Though in this case the dataset contained only metadata. But there are also model training and model inference in AI lifecycle.

Case Study: GEMA v. OpenAI (Munich district court, 11.11.2025, 42 O 14139/24):

- Background:
    - GEMA, the collective management organisation for musical copyrights, sued OpenAI, alleging ChatGPT was trained on copyrighted German song lyrics.
    - The lyrics allegedly reappeared (sometimes verbatim) in ChatGPT outputs.
- Court’s decision: Copyright infringement in both training and output. The court granted injunction, disclosure, and damages.
- Takeaways:
    - GPT-4 and GPT-4o were shown to reproduce copyrighted lyrics (“memorization”). Memorization is reproduction.
    - No justification of training via TDM exception.
    - Trainer of the models and deployer of the AI systems OpenAI was liable.
- Preliminary conclusion: Using copyrighted works for AI model training (for commercial purposes) is not permitted.

Is “training an AI model” equal to “TDM”?:

- It is complicated.

  - If “yes”, then Art. 3 CDSM will allow training an AI model using copyrighted works for research purposes without any opt-out of rightsholders.
  - If “not”, Art. 3 and Art. 4 will be irrelevant. Then, to train an AI model with copyrighted works, one needs to find other lawful ways to make it. For example, via contracts with explicit permission to use copyrighted works for training.

From the one side, there exist strong opinions against TDM=MT: “While GenAI training shares some methodological overlaps with TDM, its objectives and outputs significantly diverge. The legal and conceptual frameworks governing TDM and fair use may not seamlessly extend to generative AI, particularly given its potential to compete with and replicate the expressive elements of copyrighted works.” Dornis & Stober (2025) and Dornis (2024). From the other side, Leistner & Antoine (2025) when Art. 4 DSM is considered together with Art. 53 EU AI Act, then there is an opinion that Art. 4 DSM applies to model training.

Okay, it’s complicated, but what shall we do with model training for research purposes?

- The permissibility is not yet fully clear. The opinions differ. But:
- If you need copyrighted works for AI model training, get contractual permissions.
- If you take the risk of interpreting “AI model training” as TDM and of using the TDM exception for research purposes, make sure that:
- Requirements for the TDM exception are met (e.g., lawful access to copyrighted works)
- You don’t train a model with a third-party infrastructure.
- You train an open-weight model at a local secure hardware.
- You don’t share or archive the trained model before the legal basis for this is clarified.
- There is a need for case-by-case assessment and risk balancing.

Some of these recommendations are adapted from the talk “KI im Urheberrecht: Rechtsrahmen für Bibliothek und Wissenschaft” by Dr. Marion von Francken-Welz presented at UB Mannheim in April 2025.

License agreements and contracts:

See the full text in Agi et al. (2024).

- A blanket reference in the contract to the term “artificial intelligence” is unsuitable. One should specify concrete user acts that are to be permitted or prohibited by the contract.
- Contracts cannot effectively prohibit end users from using texts or images made available from databases as input data for AI for adaptation and transformation.

- Contracts concluded from 1 March 2018 onwards cannot effectively restrict the use of copyrighted works by TDM for scientific purposes, including the creation of internal scientific AI systems. Making reproductions available to third parties without contractual permission is only permitted under the conditions set out in section 60d (4) UrhG.
- Contracts can effectively make provision for the use of copyrighted works for TDM for non-scientific purposes.
- Contracts can provide for measures that ensure the security and integrity of the networks and databases through appropriate security precautions and set guidelines for the copies made in the context of TDM.

RAG and copyright:

There is no clear reference to RAG as a form of TDM in the existing agreements between AI developers and rightsholders.

- Static RAG may trigger more copyright-restricted acts compared to dynamic RAG. Reasons:

    - Locally hosted content often necessitates a longer retention of reproductions to enable ongoing reference, a requirement that may exceed the conditions of applicability of the CDSM Directive Article 4 TDM exception, as well as the (more strict) requirements for the applicability of the InfoSoc temporary reproduction exception.
    - By contrast, scraping the open internet for context references in dynamic RAG typically retains content only temporarily, aligning more closely with potential for application of either TDM or temporary reproduction exceptions.

More resources on copyright:

- EUIPO Copyright Knowledge Centre (https://www.euipo.europa.eu/en/copyright-knowledge-centre)
- Künstliche Intelligenz im Verlagsbereich: Häufig gestellte Fragen zu generativer KI https://www.boersenverein.de/beratung-service/recht/kuenstliche-intelligenz
- Brehm, E. (2022). Guidelines zum Text und Data Mining für Forschungszwecke in Deutschland. Brehm (2022)

Exercise 1: ❓ Quiz

Is it legal to use copyrighted works to create a dataset for training an AI model?

- Yes, always
- No, never
- Only under specific legal conditions
- It is complicated

Exercise 2: ❓ Quiz

Is it legal to use copyrighted works to train an AI model?

- Yes, always
- No, never
- Only under specific legal conditions
- It is complicated

Exercise 3: ❓ Quiz

Is it legal to use outputs of an AI system containing copyrighted works?

- Yes, always
- No, never
- Only under specific legal conditions
- It is complicated

2.d.v. *EU AI Act:*

The summary of EU AI Act is available at https://artificialintelligenceact.eu/high-level-summary.

The AI Act classifies AI according to its risk:

- Unacceptable risk is prohibited (e.g. social scoring systems and manipulative AI).
- Most of the text addresses high-risk AI systems, which are regulated.
- A smaller section handles limited risk AI systems, subject to lighter transparency obligations: developers and deployers must ensure that end-users are aware that they are interacting with AI (chatbots and deepfakes).
- Minimal risk is unregulated (including the majority of AI applications currently available on the EU single market, such as AI enabled video games and spam filters – at least in 2021; this is changing with generative AI).

EU AI Act and research:

- Art. 2 (6): „This Regulation does not apply to AI systems or AI models, including their output, specifically developed and put into service for the sole purpose of scientific research and development“ https://artificialintelligenceact.eu/article/2/
- Recital 25: „This Regulation should support innovation, should respect freedom of science, and should not undermine research and development activity. It is therefore necessary to exclude from its scope AI systems and models specifically developed and put into service for the sole purpose of scientific research and development. […] In any event, any research and development activity should be carried out in accordance with recognised ethical and professional standards for scientific research and should be conducted in accordance with applicable Union law.“

The majority of obligations fall on providers (developers) of high-risk AI systems:

- Those that intend to place on the market or put into service high-risk AI systems in the EU, regardless of whether they are based in the EU or a third country.
- And also third country providers where the high risk AI system's output is used in the EU.
- Deployers of high-risk AI systems have some obligations, though less than providers.

'provider' (developer) develops an AI system or a general-purpose AI model or has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge

'deployer' (user) uses an AI system under its authority except where the AI system is used in the course of a personal non-professional activity

Prohibited AI Systems (AI Act, Art. 5):

- AI using subliminal, manipulative, or deceptive techniques

- AI exploiting vulnerabilities related to age, disability, or socio-economic circumstances
- Biometric categorisation inferring sensitive attributes
- Social scoring systems
- Assessing the risk of an individual committing criminal offenses solely based on profiling or personality traits
- Compiling facial recognition databases by untargeted scraping of facial images
- Inferring emotions in workplaces or educational institutions
- ‘Real-time’ remote biometric identification (RBI) in publicly accessible spaces for law enforcement
    - Exceptions: missing persons; imminent threats to life or terrorism; suspects of serious crimes

High-Risk AI Systems (AI Act – Chapter III):

Scope is complex, see Art. 6.

- For example, „Education and vocational training: AI systems determining access, admission or assignment to educational and vocational training institutions at all levels. Evaluating learning outcomes, including those used to steer the student’s learning process. Assessing the appropriate level of education for an individual. Monitoring and detecting prohibited student behaviour during tests. “
- Provider requirements: risk-management system, data governance for training/validation/testing, technical documentation, built-in record-keeping, clear instructions for deployers, human-oversight capability, accuracy/robustness/cybersecurity by design, quality-management system

General purpose AI (GPAI):

All providers of GPAI models must (https://artificialintelligenceact.eu/chapter/5):

1. Draw up technical documentation, including training and testing process and evaluation results.
2. Draw up information and documentation to supply to downstream providers that intend to integrate the GPAI model into their own AI system in order that the latter understands capabilities and limitations and is enabled to comply.
3. Establish a policy to respect the Copyright Directive.
4. Publish a sufficiently detailed summary about the content used for training the GPAI model. Exception for free and open licence GPAI models: The providers only have to comply with the latter two obligations above, unless the free and open licence GPAI model is systemic.

GPAI models with systemic risks:

GPAI models present systemic risks when the cumulative amount of compute used for its training is greater than 1025 floating point operations (FLOPs).

Providers of GPAI models with systemic risk must also:

- Perform model evaluations, including conducting and documenting adversarial testing to identify and mitigate systemic risk.
- Assess and mitigate possible systemic risks, including their sources.
- Track, document and report serious incidents and possible corrective measures to the AI Office and relevant national competent authorities without undue delay.
- Ensure an adequate level of cybersecurity protection.

See https://artificialintelligenceact.eu/chapter/5.

EU AI Act and Copyright:

Specific obligations on the providers of general–purpose AI (GPAI) models (https://artificialintelligenceact.eu/article/53):

- Compliance with the TDM opt-outs expressed by copyright
- Publish 'sufficiently' detailed summaries of the training data they utilise, to facilitate copyright holders enforcing their rights holders Specific obligations on the providers of GenAI systems (https://artificialintelligenceact.eu/article/50):
- Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated.

The General-Purpose AI Code of Practice:

- The GPAI Code of Practice is a voluntary tool, prepared by independent experts in a multi-stakeholder process, designed to help industry comply with the AI Act's obligations for providers of general-purpose AI models.
- The Code of Practice has three chapters:

    1. The Transparency chapter offers a Model Documentation Form (https://ec.europa.eu/newsroom/dae/redirection/document/118118) which allows providers to easily document the information necessary to comply with the AI Act obligation to on model providers to ensure sufficient transparency.
    2. The Copyright chapter offers providers practical solutions to meet the AI Act's obligation to put in place a policy to comply with EU copyright law.
    3. The Safety and Security chapter outlines concrete state-of-the-art practices for managing systemic risks, i.e. risks from the most advanced models. Providers can rely on this chapter to comply with the AI Act obligations for providers of general-purpose AI models with systemic risk.

See https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

EU AI Act and GPAI Code of Practice: Transparency and Model Documentation Form:

- The “Transparency” chapter has one Commitment on “Documentation” (https://ec.europa.eu/newsroom/dae/redirection/document/118118)
- The Model Documentation Form contains the following parts:
    - General information
    - Model properties
    - Methods of distribution and licenses
    - (Acceptable and intended) Use
    - Training process
    - Information on the data used for training, testing, and validation
    - Computational resources (during training)
    - Energy consumption (during training and inference)

EU AI Act and GPAI Code of Practice: Copyright:

The “Copyright” chapter has one Commitment on “Copyright policy” (https://ec.europa.eu/newsroom/dae/redirection/document/118115):

- Measure 1.1 Draw up, keep up-to-date and implement a copyright policy
- Measure 1.2 Reproduce and extract only lawfully accessible copyright-protected content when crawling the World Wide Web
- Measure 1.3 Identify and comply with rights reservations when crawling the World Wide Web
- Measure 1.4 Mitigate the risk of copyright-infringing outputs
- Measure 1.5 Designate a point of contact and enable the lodging of complaints

EU AI Act and GPAI Code of Practice: Safety and Security:

The “Safety and Security” chapter has ten Commitments (https://ec.europa.eu/newsroom/dae/redirection/document/118119).

1. Commitment 1 “Safety and Security Framework”
    - Measure 1.1 Creating the Framework
    - Measure 1.2 Implementing the Framework
    - Measure 1.3 Updating the Framework
    - Measure 1.4 Framework notifications
2. Commitment 2 “Systemic risk identification”
    - Measure 2.1 Systemic risk identification process
    - Measure 2.2 Systemic risk scenarios
3. Commitment 3 “Systemic risk analysis”
    - Measure 3.1 Model-independent information
    - Measure 3.2 Model evaluations
    - Measure 3.3 Systemic risk modelling
    - Measure 3.4 Systemic risk estimation
    - Measure 3.5 Post-market monitoring
4. Commitment 4 “Systemic risk acceptance determination”

    - Measure 4.1 Systemic risk acceptance criteria and acceptance determination
    - Measure 4.2 Proceeding or not proceeding based on systemic risk acceptance determination
5. Commitment 5 “Safety mitigations”
    - Measure 5.1 Appropriate safety mitigations
6. Commitment 6 “Security mitigations”
    - Measure 6.1 Security Goal
    - Measure 6.2 Appropriate security mitigations
7. Commitment 7 “Safety and Security Model Reports”
    - Measure 7.1 Model description and behaviour
    - Measure 7.2 Reasons for proceeding
    - Measure 7.3 Documentation of systemic risk identification, analysis, and mitigation
    - Measure 7.4 External reports
    - Measure 7.5 Material changes to the systemic risk landscape
    - Measure 7.6 Model Report updates
    - Measure 7.7 Model Report notifications
8. Commitment 8 “Systemic risk responsibility allocation”
    - Measure 8.1 Definition of clear responsibilities
    - Measure 8.2 Allocation of appropriate resources
    - Measure 8.3 Promotion of a healthy risk culture
9. Commitment 9 “Serious incident reporting”
    - Measure 9.1 Methods for serious incident identification
    - Measure 9.2 Relevant information for serious incident tracking, documentation, and reporting
    - Measure 9.3 Reporting timelines
    - Measure 9.4 Retention period
10. Commitment 10 “Additional documentation and transparency”

- Measure 10.1 Additional documentation
- Measure 10.2 Public transparency

Tool: EU AI Act Compliance Checker 1:

Exercise 1: EU AI Act Compliance Checker 1

- https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker

EU AI Act Compliance Checker 2:

Exercise 2: EU AI Act Compliance Checker 2

- https://ai-act-service-desk.ec.europa.eu/en/eu-ai-act-compliance-checker

Exercise 3: ❓ Quiz

Can a researcher train an AI model with systemic risk and openly share it?

- Yes
- No
- Yes, but with obligations

2.e. AI risks, AI safety, and AI security

According to Lin et al. (2025), AI safety and AI security differ in the following way:

- AI Safety: risks from accidental or unintended behaviors.
- AI Security: risks from intentional or adversarial actions by malicious actors.

2.e.i. *Main documents on AI safety and AI security:*

- The International AI Safety Report is the world's first comprehensive review of the latest science on the capabilities and risks of general-purpose AI systems (∼ 300 pages):
  - https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025
  - https://internationalaisafetyreport.org/publication/first-key-update-capabilities-and-risk-implications
  - https://internationalaisafetyreport.org/publication/second-key-update-technical-safeguards-and-risk-management
- OWASP AI Exchange is a practical resource on AI security and privacy (>200 pages)
  - https://owaspai.org
- Federal Office for Information Security issues studies, catalogues, and checklists on AI: https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/kuenstliche-intelligenz_node.html

2.e.ii. *AI Safety risks from the AI Safety report:*

Selected risks from the AI safety report relevant in research context:

1. Risks from malicious use
   - Harm to individuals through fake content. AI-generated fake content can be used to manipulate governance processes, sabotage collaborations, or personally target researchers and undermine trust in science.
   - Manipulation of public opinion. AI-generated fake content and narrative manipulation can directly target researchers, research projects, research ideas, and public trust in science.
   - Cyber offence. AI-assisted cyber attacks may affect research infrastructures, collaborations, and sensitive scientific outputs.
   - Biological and chemical attacks.
2. Risks from malfunctions
   - Reliability issues. This can lead violations of research ethics, research integrity, and research governance due to irresponsible use of AI systems having reliability issues.

  - Bias. Biased data, models, and AI systems can lead to discriminatory or misleading results violating research ethics principles. Researchers have an ethical duty to identify, mitigate, and transparently report bias.
3. Systemic risks
  - Risks to the environment. Key ethical question for research projects: Is the environmental cost of using an AI system justified by the expected scientific and societal benefit?
  - Risks to privacy. This can lead to violations of research ethics, research integrity, and research governance.
  - Risks of copyright infringement. This can lead to violations of research ethics, research integrity, and research governance.
4. Impact of open-weight general-purpose AI models on AI risks.

Risks from open-weight general-purpose AI models on AI risks:

The AI safety report has the following points:

- Risks posed by open-weight models are largely related to enabling malicious or misguided use, because general-purpose AI models are dual-use:
- Safeguards against misuse are easier to remove for open models
- Model vulnerabilities found in open models can also expose vulnerabilities in closed models
- Once model weights are available for public download, there is no way back
- There are risk mitigation approaches for open-weight models throughout the AI lifecycle. The most robust risk mitigation strategies will aim to address potential issues at every stage from data collection to post-release measures such as vulnerability disclosure.

From the other side:

- Security by obscurity is not recommended by NIST: “System security should not depend on the secrecy of the implementation or its components.”
- The Kerckhoffs’s principle: “A system should be secure, even if everything about the system is public knowledge.”
- “In modern computer security, a hard-fought broad-based consensus has been established: Despite the intuitive idea that hiding a system should protect it, transparency is often more beneficial for protection. The consensus on this general principle is broad, though perspectives on how to implement the principle in specific contexts can be more varied.” Hall et al. (2025)
- In research: Open Science + Safety + Security = As open as possible, as closed as necessary.

Exercise 1: gpt-oss-safeguard

- Open-weight safety reasoning models
- Trained to reason about safety
- Support custom safety policies

- Configurable reasoning effort
- Apache 2.0 license
- Test it: https://huggingface.co/spaces/openai/gpt-oss-safeguard-20b
- Sample policy: example_policies/spam/policy.txt

Exercise 2: ❓ Quiz

A researcher is an open science enthusiast and shares code, data, and model weights of an in-house developed AI system openly and in accordance with FAIR principles. How can dual-use risks be prevented?

- Open science automatically prevents misuse
- Share only model weights, but not data
- Apply risk mitigation measures to every stage of AI lifecycle
- Avoid publishing AI research entirely

2.e.iii. *AI Security Exchange:*

Threat model: Types of threats:

Three types of threats (https://owaspai.org/goto/threatsoverview):

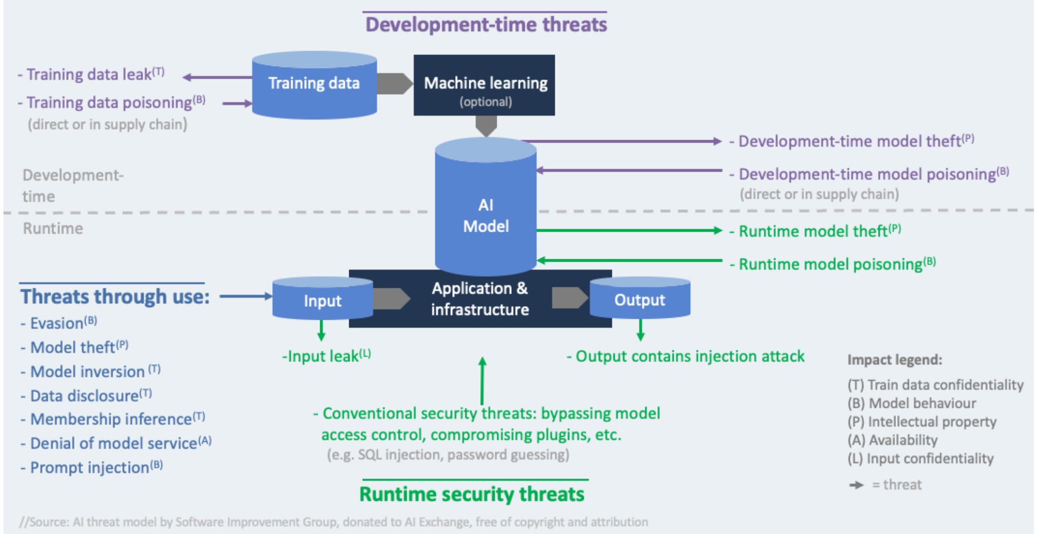
1. threats during development-time: when data is obtained and prepared, and the model is trained/obtained. Example: data poisoning (injecting bad data into the training data)
2. threats through using the model: through inference; providing input and getting the output. Examples:
   - direct prompt injection (malicious prompt into the user interface),
   - indirect prompt injection (malicious prompt is embedded in external content) or
   - evasion (hidden malicious instructions via obfuscation, encoding, hidden text, and payload splitting)
3. other threats to the system during runtime: in operation - not through inference. Example: stealing model input

Threat model: Impacts:

6 types of impacts that align with three types of attacker goals (disclose, deceive and disrupt):

1. disclose: hurt confidentiality of train/test data
2. disclose: hurt confidentiality of model Intellectual property (the model parameters or the process and data that led to them)
3. disclose: hurt confidentiality of input data
4. deceive: hurt integrity of model behaviour (the model is manipulated to behave in an unwanted way and consequentially, deceive users)

5. disrupt: hurt availability of the model (the model either doesn't work or behaves in an unwanted way - not to deceive users but to disrupt normal operations)
6. disrupt/disclose: confidentiality, integrity, and availability of non AI-specific assets


Development-time threats
- Training data leak(T)
- Training data poisoning(B)
(direct or in supply chain)
Training data
Machine learning
(optional)
- Development-time model theft(P)
- Development-time model poisoning(B)
(direct or in supply chain)
Development-
time
Runtime
AI
Model
- Runtime model theft(P)
- Runtime model poisoning(B)
Threats through use:
- Evasion(B)
- Model theft(P)
- Model inversion (T)
- Data disclosure(T)
- Membership inference(T)
- Denial of model service(A)
- Prompt injection(B)
Input
Application &
infrastructure
Output
-Input leak(L)
- Output contains injection attack
- Conventional security threats: bypassing model
access control, compromising plugins, etc.
(e.g. SQL injection, password guessing)
Runtime security threats
Impact legend:
(T) Train data confidentiality
(B) Model behaviour
(P) Intellectual property
(A) Availability
(L) Input confidentiality
= threat

//Source: AI threat model by Software Improvement Group, donated to AI Exchange, free of copyright and attribution

Threats to agentic AI:

Threats:

- Hallucinations and prompt injections can change commands or even escalate privileges.
- Leak of sensitive data due to the „lethal trifecta“:
  - Data: Control of the attacker of data that find its way into an LLM at some point in the session of a user that has the desired access, to perform indirect prompt injection
  - Access: Access of that LLM or connected agents to sensitive data
  - Send: The ability of that LLM or connected agents to initiate sending out data to the attacker

Exercise 3: ❓ Quiz

Which security risk is introduced when researchers use an AI system (e.g., an agentic research assistant or coding assistant such as Cursor) that has unrestricted access to local files, code repositories, and credentials?

- Reduced model accuracy
- Increased energy consumption
- Unauthorized data access and data exchange with third parties
- Loss of explainability

Exercise 4: ❓ Quiz

Which of the following is a core AI security concern (as opposed to AI safety)?

- Bias against minority groups
- Misleading scientific conclusions
- Intentional misuse or exploitation of AI systems by adversaries
- Lack of reproducibility

2.f. Risk management

ISO 31000 is an international standard providing principles, a framework, and a process for organizations to manage risks

Specific AI risk management documents:

- ISO/IEC 23894:2023 is an international standard that provides guidance on managing risks associated with AI for organizations that develop, produce, deploy, or use AI.
- The NIST AI Risk Management Framework (AI RMF) is a framework to better manage risks to individuals, organizations, and society associated with AI (Tabassi (2023)).
- Chapter 3 of the AI Safety Report provides overview of AI risk management approaches (https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025#3-technical-approaches-to-risk-management)
- AI Exchange on AI Security refers to the Risk Management Framework of ISO/IEC 23894:2023 (https://owaspai.org/goto/riskanalysis/).
- Guidance for Risk Management of Artificial Intelligence systems by European Data Protection Supervisor (https://www.edps.europa.eu/system/files/2025-11/2025-11-11_ai_risks_management_guidance_en.pdf)

2.f.i. *Selected risk management mechanisms and practices from the AI safety report:*

- Risk Identification: Risk Taxonomy, Threat Modelling
- Risk Assessment:
  - Impact Assessment (e.g., Fundamental Rights Impact Assessment for High-Risk AI Systems due to the EU AI Act https://artificialintelligenceact.eu/article/27, Data Protection Impact Assessment due to GDPR https://gdpr-info.eu/art-35-gdpr). Question: What's about copyright & Co?
  - Audit (https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-auditing_en)
  - Red-Teaming (deliberately attacking your own AI systems in order to identify vulnerabilities)
  - Benchmarks
  - Model Evaluation
  - Safety Analysis
- Risk Evaluation: Risk Tolerance, Risk Thresholds
- Risk Mitigation:
  - Safety by Design (remember Chapter 3 "Safety and Security" of the Code of Conduct EU AI Act)
  - Defense in Depth (implementing multiple independent and overlapping layers of defense, multiple preventive measures)
- Risk Governance:

  - Documentation (model cards, system cards, etc. Remember Chapter 1 “Transparency” of the Code of Conduct EU AI Act)
  - Risk Register
  - Incident Reporting
  - Risk Management Frameworks

2.f.ii. *Risk management framework:*

- Identifying Risks: Recognizing potential risks that could impact the organization.
- Evaluating Risks by Estimating Likelihood and Impact: To determine the severity of a risk, it is necessary to assess the probability of the risk occurring and evaluating the potential consequences should the risk materialize.
- Deciding What to Do (Risk Treatment): Choosing an appropriate strategy to address the risk. These strategies include:
  - Risk Mitigation,
  - Transfer,
  - Avoidance,
  - Acceptance.
- Risk Communication and Monitoring: Regularly sharing risk information with stakeholders to ensure awareness and continuous support for risk management activities. Ensuring effective Risk Treatments are applied. This requires a Risk Register, a comprehensive list of risks and their attributes (e.g. severity, treatment plan, ownership, status, etc).

2.f.iii. *Risk identification via risk taxonomy: MIT AI Risk Repository:*

Slattery et al. (2024)

Exercise 1: ❓ Quiz

Which of the following best describes risk management of AI use in research projects?

- Avoiding AI use whenever risks are identified
- Identifying, evaluating, treating, and monitoring risks throughout the research lifecycle
- Delegating all responsibility to IT departments
- Focusing only on legal compliance after deployment

3. 2. Practice

In the practical part we continue with:

- Risk management framework adapted to the trinity of risks (ethical, integrity, and governance)
- Hands-On: Use cases of AI and risk management across the research lifecycle. Plan & design, collect & create, analyze & collaborate, evaluate & archive, share & disseminate, access & reuse.
- AI policies and checklists for research groups and research projects

3.a. Risk management in research projects

3.a.i. *Defense in depth for your research (project):*

Defense in depth is a risk-mitigation approach that relies on multiple, overlapping, and complementary layers of safeguards, so that if one control fails, others remain effective.

Key assumption: any single mitigation measure can fail.

Dimensions of defense in research (projects) for an AI use case:

- People: individual researchers, research teams, project partners, and participants
- Infrastructure/Technology: hardware, software, data, models, networks, and the entire supply chain
- Processes: research process including research ethics, research integrity, and research governance

3.a.ii. *Trinity of risks related to AI use in research:*

Using AI in research (projects) introduces the following risks:

- Ethical: intentional and unintentional harms to participants, communities, society, and the environment (privacy and data protection, copyright and other intellectual rights, confidentiality, bias, discrimination, malicious use, and dual use)
- Integrity: intentional and unintentional harms to the truth and scientific quality including scientific misconduct, questionable research practices, and irresponsible use of AI systems
- Governance: intentional and unintentional breaches of the laws and regulations including GDPR, copyright, EU AI Act, dual-use and export control regulations, and contractual requirements

3.a.iii. *Practical risk management:*

There is an AI use case in a research project.

- It contains three dimensions: people (users, developers, co-authors, etc.), technology (AI system and alternatives), and processes (ethics, integrity, and governance).
- Apply the risk management framework adapted to the trinity of risks in research projects to your AI use case. [see the next slide]

3.a.iv. *Risk management framework adapted to the trinity of risks in research (projects):*

Let’s adapt the Risk management framework (https://owaspai.org/goto/riskanalysis/) to the trinity of risks:

- Identifying Risks: Recognizing potential risks related to research ethics, research integrity, and research governance.
- Evaluating Risks: Subjectively: low, medium, high.

- Risk Treatment: Choosing an appropriate strategy to address the risk: Risk Mitigation, Transfer (to another party), Avoidance (eliminating the source), or Acceptance.
- Risk Communication and Monitoring: Regularly sharing risk information with stakeholders. Creating a Risk Register, a list of risks and their attributes (e.g. severity, treatment plan, ownership, status, etc).

> **Note**
>
> *Any mitigation measure can fail. Apply defense in depth to all three dimensions (people, technology, and processes) in Risk Mitigation.

3.a.v. *People (Researchers, Teams, and Partners):*

General mitigation measures involving people and human factor

- Competence & awareness
  - Define minimum AI literacy expectations for project members (what AI can/cannot do).
  - Ensure researchers understand the trinity or risks. Treat AI literacy as continuous.
- Clear responsibility
  - Explicitly state: humans remain responsible for scientific outputs, not AI systems.
  - Assign AI-use responsibility at project level (e.g., PI or work-package lead).
- Disclosure & transparency
  - Agree on when and how AI use must be disclosed
  - Encourage a culture where declaring AI use is normal, not penalised.
- Collaboration safeguards
  - Align expectations on AI use with project partners, co-authors, and student assistants
  - Explicitly clarify AI rules for junior researchers and external collaborators.

3.a.vi. *Technology & Infrastructure: AI Systems:*

General mitigation measures involving technology and infrastructure

- Tool selection
  - Prefer AI systems (including models) that provide compliance documentation (trust portals and certifications)
- Access control
  - Apply least-privilege principles to AI systems (e.g., limit file system access)
  - Never grant agentic AI unrestricted access to personal, sensitive, and confidential data

- Data protection
  - Don’t input personal or sensitive data into third-part AI systems. Use fully local, self-managed AI systems
  - Use pseudonymisation, anonymisation, or synthetic data where possible.
- Model limitations
  - Any AI model has limitations. Check them.
- Technical safeguards
  - Ensure appropriate technical and organisational measures

3.a.vii. *Processes: Ethics, Integrity, and Governance:*

General mitigation measures for three processes in research

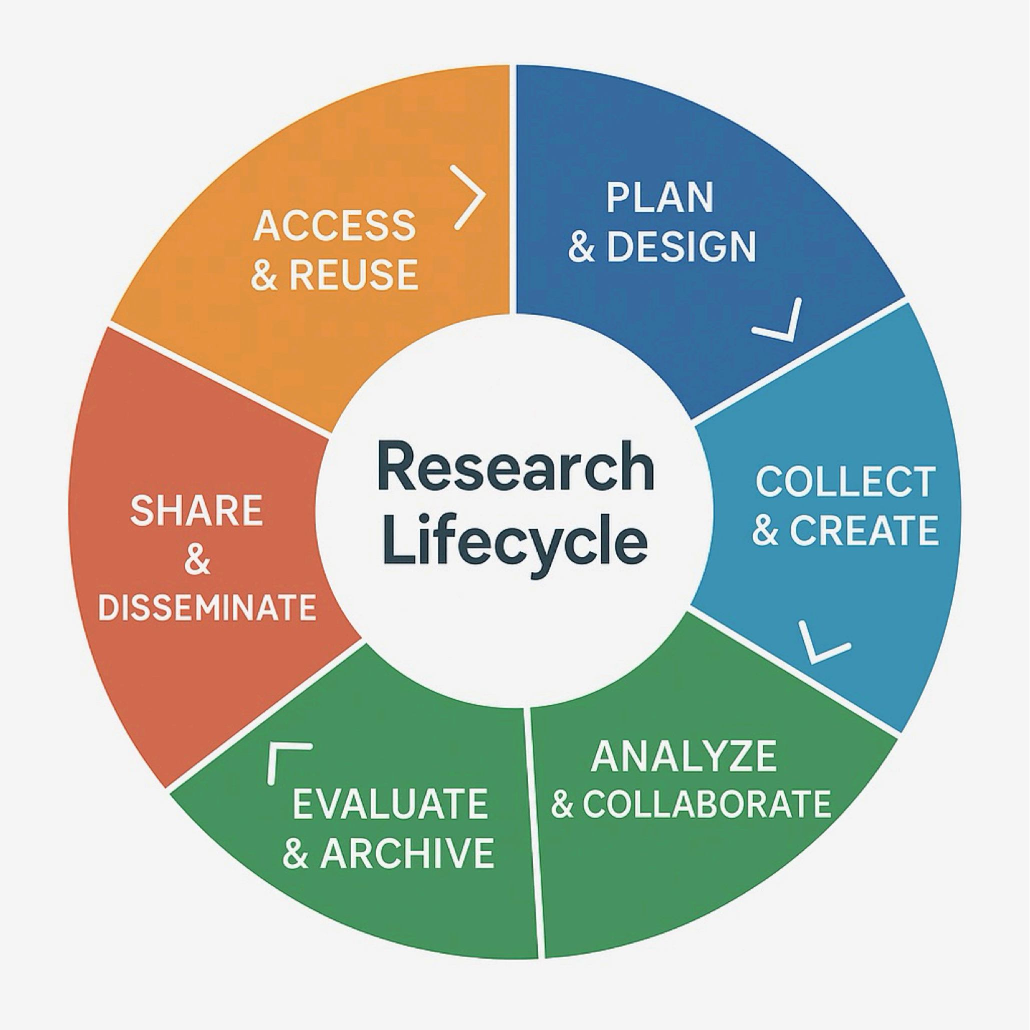
- Research ethics:
  - Integrate AI use into ethics self-assessment, even if formal ethics approval is not required.
  - Get support from ethics committees
- Research integrity:
  - Require AI-use documentation
  - Preserve input and output data, models, and ensure reproducibility/replicability of your research
  - Define red lines: no fabrication, falsification, and plagiarism
- Research governance:
  - Apply legal and regulatory checklists
  - Conduct legal pre-checks
  - Get institutional support: DPO, legal team, RDM-team, IT- and IT-security-teams, etc.

3.a.viii. *Mapping the trinity of risks and existing guidelines and recommendations on AI use in research:*

Trinity of risks	EU Guidelines on the responsible use of generative AI in research	Helmholtz “Recommendations for the use of artificial intelligence”
Ethical risks	· Privacy, confidentiality, and IP rights · Bias and prejudices due to training data	· Privacy and confidentiality
Integrity risks	· Responsibility for scientific output · Transparent AI use · Continuous AI literacy	· Sensitive activities impacting others · (Scientific) information integrity

Trinity of risks	EU Guidelines on the responsible use of generative AI in research	Helmholtz "Recommendations for the use of artificial intelligence"
Governance risks	· National, EU & international legislation	· AI-related regulation · Copyright and IP rights · Privacy and confidentiality

3.b. Hands-On: Use cases of AI and risk management across the research lifecycle



3.b.i. *Hands-On Exercise (20 min): For every stage in the research lifecycle:*

Risk management in every stage

- Pick: one AI use case
- Identify:
  - Multiple ethical risks
  - Multiple integrity risks
  - Multiple governance risks
- Rate severity: low, medium, and high
- Choose multiple treatment strategies: Risk Mitigation, Transfer (to another party), Avoidance (eliminating the source), or Acceptance. In Risk mitigation apply defense in depth to all three dimensions (people, technology, and processes).
- Create: a risk register for your AI use case.

3.c. *AI policies and checklists for research groups and research projects*

3.c.i. *Why research groups and projects need their own AI policies and checklists:*

Institutional AI policies and checklists are necessary but not sufficient. Research groups and projects may need context-specific rules that reflect their data, methods, risks, and responsibilities.

Reasons:

- AI risks are discipline-specific and use-case-specific
- In international research projects even the regulations for research ethics, research integrity, and research governance may differ

3.c.ii. *Purpose of AI Policies & Checklists:*

AI policies and checklists help research groups and research projects:

- prevent ethical, integrity, and governance breaches
- clarify responsibilities
- ensure transparency and reproducibility
- support safe and secure AI use

3.c.iii. *What is AI policy for a research group or project?:*

An AI policy is:

- A living document
- A risk management tool and risk register
- A shared agreement within the group/project
- A bridge between ethics, integrity, and governance

3.c.iv. *Scope, responsibilities, and minimal compliance:*

- Which AI systems are allowed?
- For which tasks?
- Who is responsible for:
  - AI selection
  - risk management
  - documentation
  - incident reporting
- Minimal compliance: GDPR, Copyright, EU AI Act, export control, cybersecurity, etc.

3.c.v. *People-Checklist:*

Researchers & Team Members

- Basic AI literacy for all team members
- Awareness of ethical risks (harms to participants, society, and the environment)
- Understanding of research integrity rules for AI use

- Clear rules for disclosure of AI use
- No “shadow AI” or undocumented tool usage

Collaboration & Culture

- AI use discussed openly in the team
- Agreement on what counts as acceptable AI assistance
- Clear expectations for students, HiWis, and PhDs
- Special care for interdisciplinary & international projects

3.c.vi. *Technology-Checklist:*

AI Systems & Tools

- Approved AI tools list (local, cloud, or hybrid)
- Trust portals & compliance documentation checked (cloud and hybrid)
- Data residency & logging behavior understood
- No personal or sensitive data in public AI systems

Security Controls

- Least-privilege access to data, code, networks, and credentials
- No unrestricted agentic AI access to local files
- Versioning of models, prompts, and outputs
- Monitoring for data leakage and misuse

Open-Weight Models

- Risk assessment before downloading model weights
- No public release without governance review
- Vulnerability reporting plan in place

3.c.vii. *Processes-Checklist:*

Ethics (Is it acceptable? Is it responsible towards participants, society, and the environment?)

- Could AI use distort interpretation or fairness?
- Could vulnerable groups be affected indirectly?
- Are societal or environmental impacts considered?

Integrity (Is it good science? Is it responsible towards the truth?)

- AI use documented in methods sections
- Original sources preserved and cited
- Human judgment remains central
- Reproducibility ensured despite AI variability

Governance (Is it compliant?)

- DPIA completed if required

- Licenses and terms respected
- Ethics committee informed if risk profile changes
- Archiving and sharing rules defined

3.c.viii. *Last but not least:*

Rewrite your group or project AI policy as “10 simple rules for using AI” in your field.

4. Conclusions

AI systems and AI models increasingly support all stages of research projects across the entire research lifecycle. They are used for literature review, research design, idea generation, data collection and analysis, content generation, coding, writing, archiving, sharing, and dissemination. In this workbook, we addressed the safe and secure use of AI in research projects through the prism of the trinity of good research: research ethics, research integrity, and research governance.

In the Introduction, we began with key definitions of AI systems and AI models, AI safety and AI security, and risk management. We highlighted that, before using any AI system, researchers should consider the safety and security of all its components: hardware, software, data, models, and networks.

In Part 1 (Theory), we showed that freedom of research is exercised within ethical, integrity, and governance constraints, and that AI both amplifies existing risks and introduces new ones. AI safety and AI security correspond to risks that arise unintentionally as well as deliberately. These AI-related risks intersect with research ethics, research integrity, and research governance, each of which addresses different responsibilities. Research ethics concerns responsibility towards participants, society, and the environment. Research integrity focuses on responsibility towards scientific truth and, consequently, the behaviour of researchers. Research governance addresses responsibility for legal and regulatory compliance. We also reviewed existing AI risk management approaches and introduced a general risk management framework.

In Part 2 (Practice), we adapted this risk management framework to the trinity of risks in research: ethical, integrity-related, and governance-related. By identifying risks, evaluating their likelihood and impact, and selecting appropriate treatment strategies (mitigation, transfer, avoidance, or acceptance), researchers can make informed decisions about AI use. We then discussed AI use cases across the entire research lifecycle and demonstrated that AI-related risks occur at every stage of research. Because any single safeguard can fail, so-called defense in depth across people, technology, and processes is essential as a mitigation approach. We also provided general mitigation strategies covering all three dimensions.

This leads to a key practical message: research groups and research projects need their own AI policies and checklists. Institutional rules are necessary but insufficient, as they cannot fully address domain-specific and project-specific contexts. Group-level AI policies function as living agreements that clarify responsibilities, prevent undocumented “shadow AI”, support transparency and accountability, and protect researchers. To support this, checklists for people, technology, and processes are provided.

Safe, secure, and responsible use of AI in research is a continuous process. Start discussing these topics within your research group and with project partners. Raise awareness, identify and mitigate risks, and revisit decisions as projects evolve. Responsible AI use is a team effort. Research ethics, research integrity, and research governance are not box-ticking exercises. They are quality instruments that enable responsible and excellent research in the age of AI.

References

Agi, C., Beurskens, M., Francken-Welz, M. von, Ludwig, J., Mittermaier, B., & Pampel, H. (2024). *Arrangements on artificial intelligence in licence agreements.* https://doi.org/10.5281/ZENODO.13837688

ALLEA. (2023,). *The European Code of Conduct for Research Integrity – Revised Edition 2023.* ALLEA. https://doi.org/10.26356/ECOC

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., … Zeng, Y. (2025,). *International AI Safety Report.* arXiv. https://doi.org/10.48550/ARXIV.2501.17805

Brehm, E. (2022). *Guidelines zum Text und Data Mining für Forschungszwecke in Deutschland* [Techreport]. https://doi.org/10.34657/9388

Community, T. T. W., & Scriberia. (2024,). *Illustrations from The Turing Way: Shared under CC-BY 4.0 for reuse.* Zenodo. https://doi.org/10.5281/zenodo.13882307

Deutsche Forschungsgemeinschaft. (2025). *Guidelines for Safeguarding Good Research Practice. Code of Conduct.* https://doi.org/10.5281/ZENODO.14281892

Dornis, T. W. (2024). *The Training Of Generative Ai Is Not Text And Data Mining.* https://doi.org/10.2139/ssrn.4993782

Dornis, T. W., & Stober, S. (2025,). *Generative AI Training and Copyright Law.* arXiv. https://doi.org/10.48550/ARXIV.2502.15858

Enrico Glerean. (2025). *Fundamentals of Secure AI Systems with Personal Data* [Training curriculum on AI and data protection]. https://www.edpb.europa.eu/system/files/2025-06/spe-training-on-ai-and-data-protection-technical_en.pdf

European Union Intellectual Property Office. (2025). *The development of generative artificial intelligence from a copyright perspective.* Publications Office. https://doi.org/10.2814/3893780

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities.* Oxford University PressOxford. https://doi.org/10.1093/oso/9780198883098.001.0001

Frisch, K. (2025). *FAQ Künstliche Intelligenz und gute wissenschaftliche Praxis - Version 2.* https://doi.org/10.5281/ZENODO.17349995

Glynn, A. (2024,). *Academ-AI: documenting the undisclosed use of generative artificial intelligence in academic publishing.* arXiv. https://doi.org/10.48550/ARXIV.2411.15218

Hall, P., Mundahl, O., & Park, S. (2025). The Pitfalls of ``Security by Obscurity'' and What They Mean for Transparent AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(27), 28042–28051. https://doi.org/10.1609/aaai.v39i27.35022

He, H. (2025). *Defeating Nondeterminism in LLM Inference.* Thinking Machines Lab: Connectionism. https://doi.org/10.64434/tml.20250910

Iacino, G., Kamocki, P., & Leinen, P. (2023). *Assessment of the Impact of the DSM-Directive on Text+.* https://doi.org/10.5281/ZENODO.12759960

Kolstoe, S. (2023). *Defining the Spectrum of Questionable Research Practices (QRPs).* UK Research Integrity Office. https://doi.org/10.37672/ukrio.2023.02.qrps

Kolstoe, S. E., & Pugh, J. (2024). The trinity of good research: Distinguishing between research integrity, ethics, and governance. *Accountability in Research*, *31*(8), 1222–1241. https://doi.org/10.1080/08989621.2023.2239712

Koster, M., Illyes, G., Zeller, H., & Sassman, L. (2022). *Robots Exclusion Protocol.* RFC Editor. https://doi.org/10.17487/rfc9309

Leistner, M., & Antoine, L. (2025). TDM and AI Training in the European Union – From ‘ LAION ’ to Possible Ways Ahead?. *GRUR International*, *74*(11), 1027–1044. https://doi.org/10.1093/grurint/ikaf114

Lin, Z., Sun, H., & Shroff, N. (2025,). *AI Safety vs. AI Security: Demystifying the Distinction and Boundaries.* arXiv. https://doi.org/10.48550/ARXIV.2506.18932

Meho, L. I. (2025). Gaming the metrics: bibliometric anomalies in global university rankings and the research integrity risk index (RI2). *Scientometrics*, *130*(11), 6683–6726. https://doi.org/10.1007/s11192-025-05480-2

Shigapov, R. (2025b). *Questionable Practices in the use of AI.* https://doi.org/10.5281/ZENODO.17349510

Shigapov, R. (2025a,). *Safe and secure use of AI in research projects*. Zenodo. https://doi.org/10.5281/ZENODO.17940942

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024,). *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence*. arXiv. https://doi.org/10.48550/ARXIV.2408.12622

Suchikova, Y., Tsybuliak, N., Silva, J. A. Teixeira da, & Nazarovets, S. (2025). GAIDeT (Generative AI Delegation Taxonomy): A taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. *Accountability in Research*, 1–27. https://doi.org/10.1080/08989621.2025.2544331

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards. https://doi.org/10.6028/nist.ai.100-1

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2022,). *Recommendation on the Ethics of Artificial Intelligence*. UNESCO. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Yuan, J., Li, H., Ding, X., Xie, W., Li, Y.-J., Zhao, W., Wan, K., Shi, J., Hu, X., & Liu, Z. (2025,). *Understanding and Mitigating Numerical Sources of Nondeterminism in LLM Inference*. arXiv. https://doi.org/10.48550/ARXIV.2506.09501