# Trustworthy AI and its Connection to Reproducibility

INVITED TALK

Dominik Kowald
dkowald@know-center.at

KFU Graz
April 2024, Graz, Austria

# Agenda

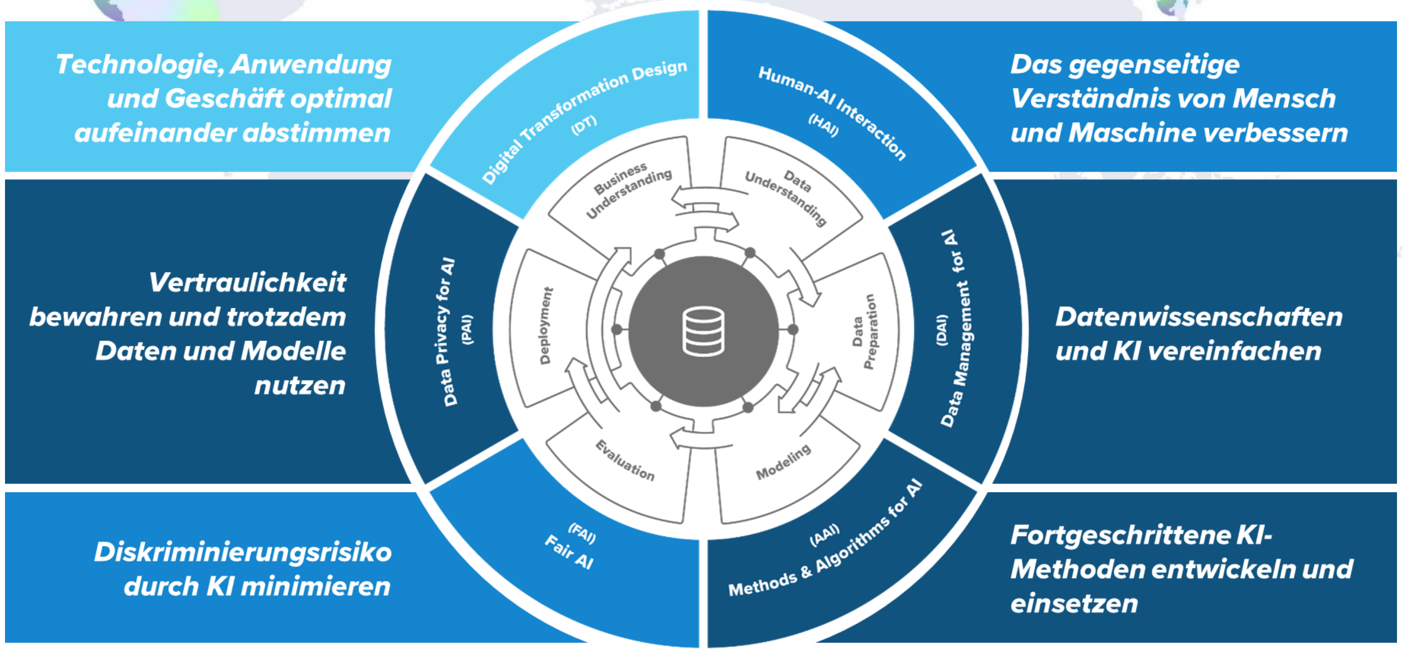
1. Intro + FAIR-AI research area
2. Trustworthy AI: what does it mean and how can we validate it?
3. Reproducibility in AI and ML-driven research
    1. Definition
    2. Barriers
    3. Drivers
4. Conclusion and some suggestions

KNOW Center

# Know-Center GmbH

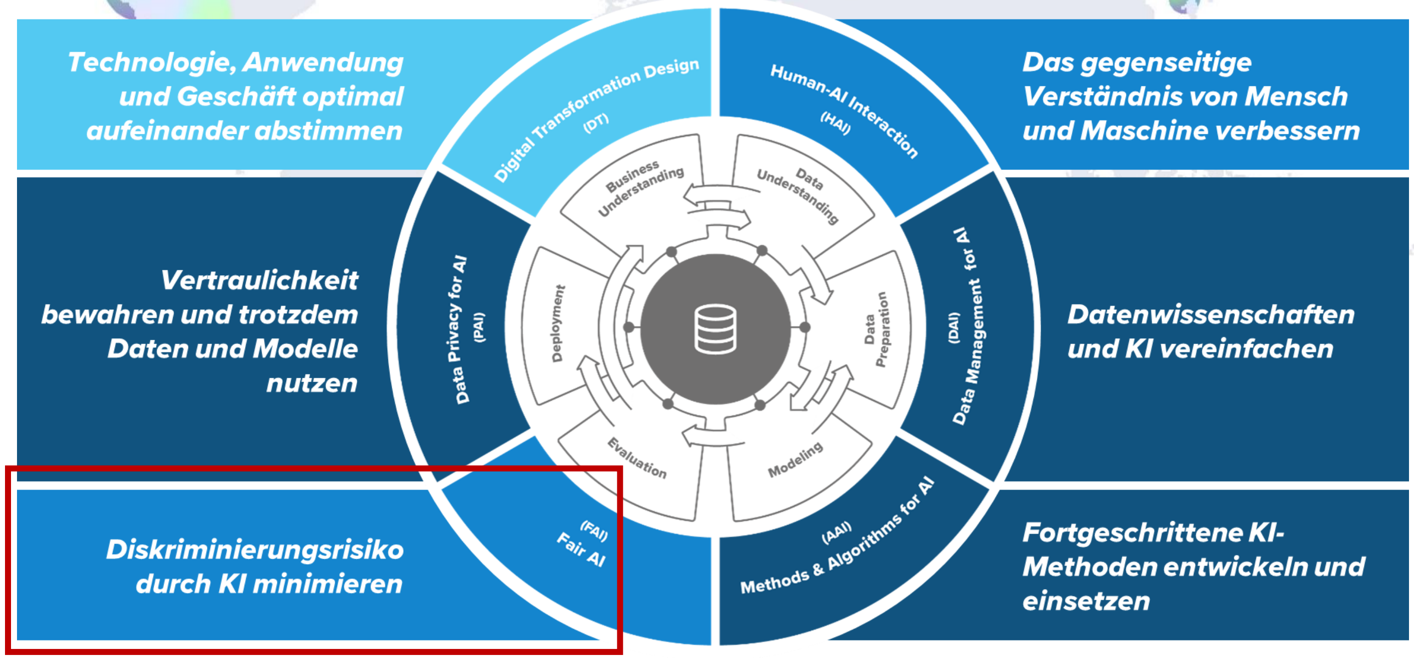
- COMET center
- Founded in 2001
- 100+ employees
- TU Graz, Campus Inffeldgasse

*„We research, develop, and provide (consulting) services along the data value chain on the topic of **trustworthy AI and data science**”*



# Know-Center GmbH

- COMET center
- Founded in 2001
- 100+ employees
- TU Graz, Campus Inffeldgasse



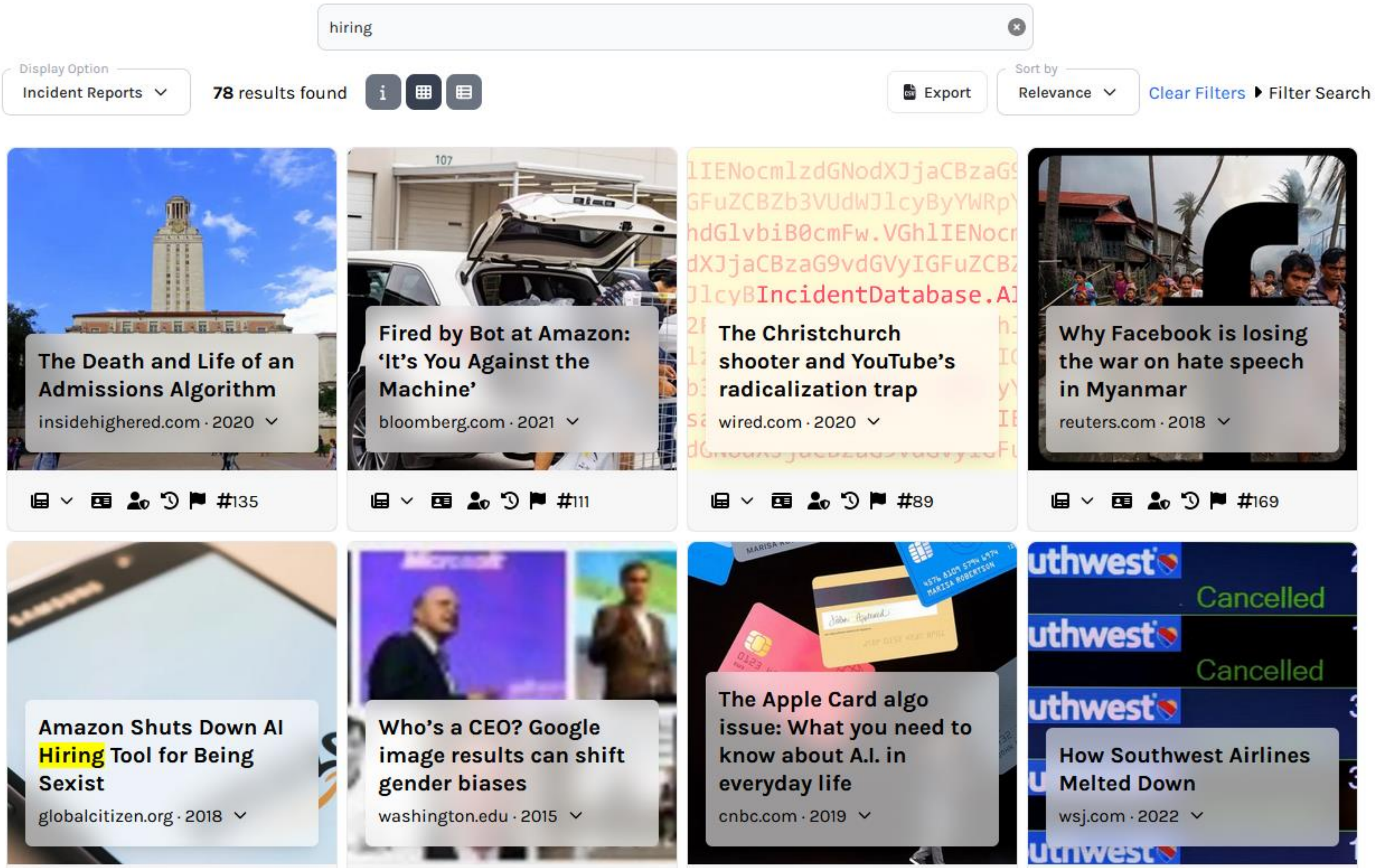

*„We research, develop, and provide (consulting) services along the data value chain on the topic of **trustworthy AI and data science**"*

# Trustworthy AI:
# what does it mean and how can we validate it?

# AI – what could go wrong?

[https://incidentdatabase.ai/]

- AI incidents database
  - > 2000 incidents
  - ~ 80 for hiring

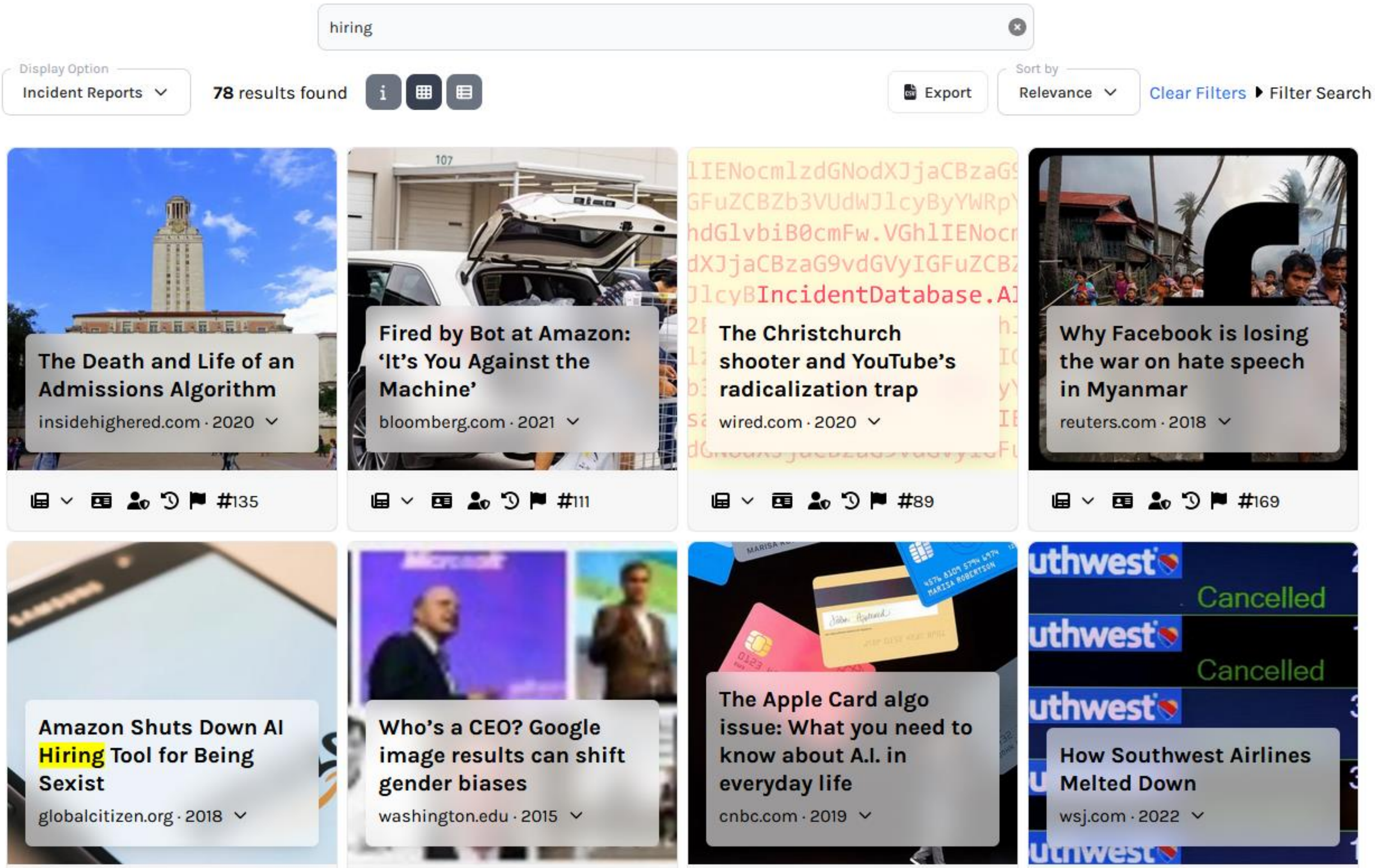


KNOW Center

# AI – what could go wrong?

[https://incidentdatabase.ai/]

- AI incidents database
  - > 2000 incidents
  - ~ 80 for hiring

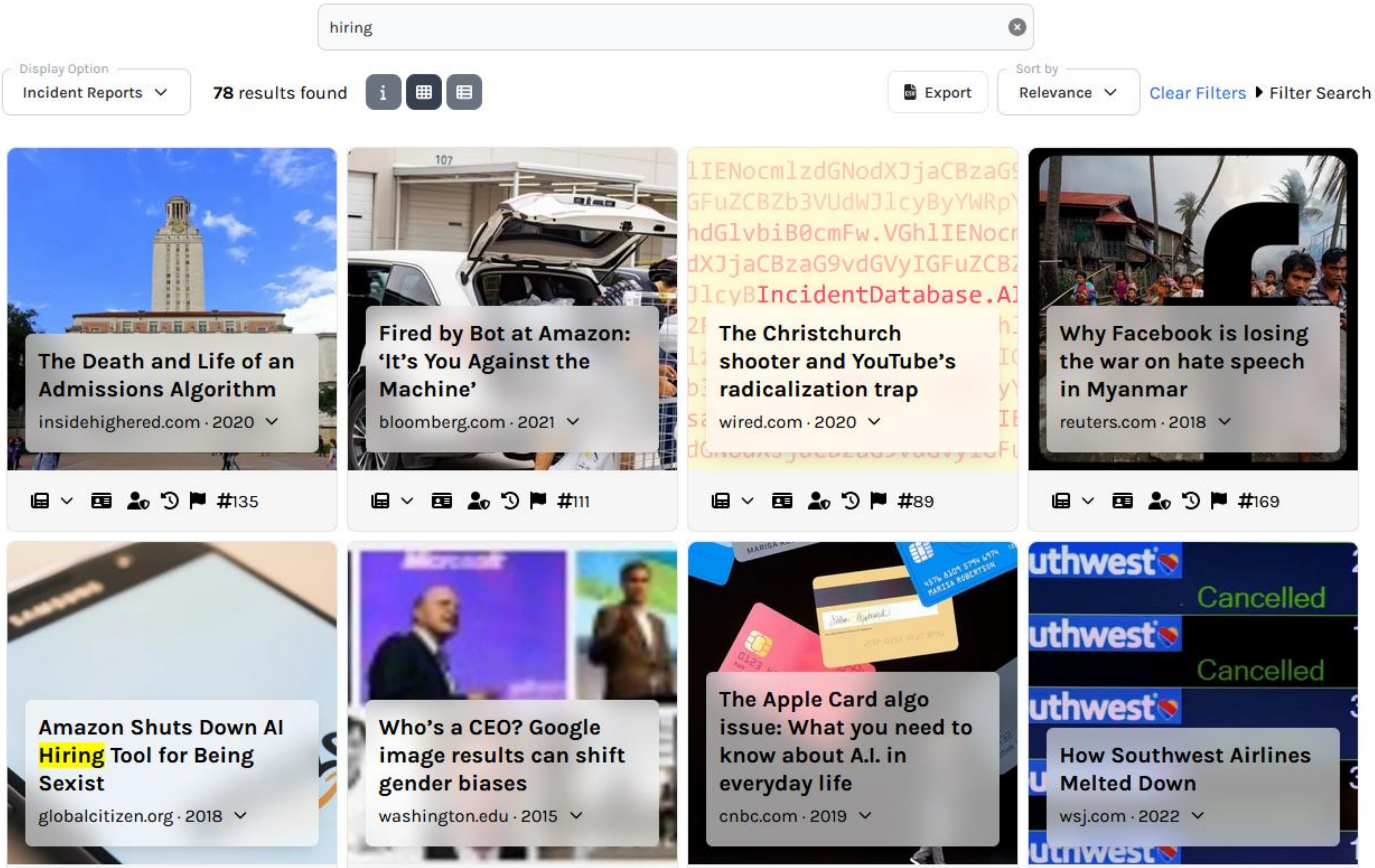


# AI – what could go wrong?

[https://incidentdatabase.ai/]

- AI incidents database
  - > 2000 incidents
  - ~ 80 for hiring

- Why?
  - Historic biases in data
  - Unclear definitions of trustworthiness
  - AI systems are continually learning systems



# AI – what could go wrong?

- Other examples
  - Data is leaked (privacy)
  - AI models are tricked (robustness)
  - AI models not usable in health care due to lack of explainability (transparency)

  …

  - ML/AI-driven research not reproducible …

# EU AI-Act (provisional time-line, starting 2024)

## Legal framework

Mid 2024
Entry into force

After 6 months
(early 2025)
Member states shall phase out prohibited systems (e.g., military)

After 12 months
(mid 2025)
Provisions on Foundation Models apply

After 24 months
(mid 2026)
Requirements for High Risk Systems apply

After 36 months
(mid 2027)
Requirements for all risk systems apply

# EU AI-Act (provisional time-line, starting 2024)

## Legal framework

Mid 2024
Entry into force

After 6 months
(early 2025)
Member states shall phase out prohibited systems (e.g., military)

After 12 months
(mid 2025)
Provisions on Foundation Models apply

After 24 months
(mid 2026)
Requirements for High Risk Systems apply

After 36 months
(mid 2027)
Requirements for all risk systems apply

# EU AI-Act (provisional time-line, starting 2024)

## Legal framework

Mid 2024
Entry into force

After 6 months
(early 2025)
Member states shall phase out prohibited systems (e.g., military)

After 12 months
(mid 2025)
Provisions on Foundation Models apply

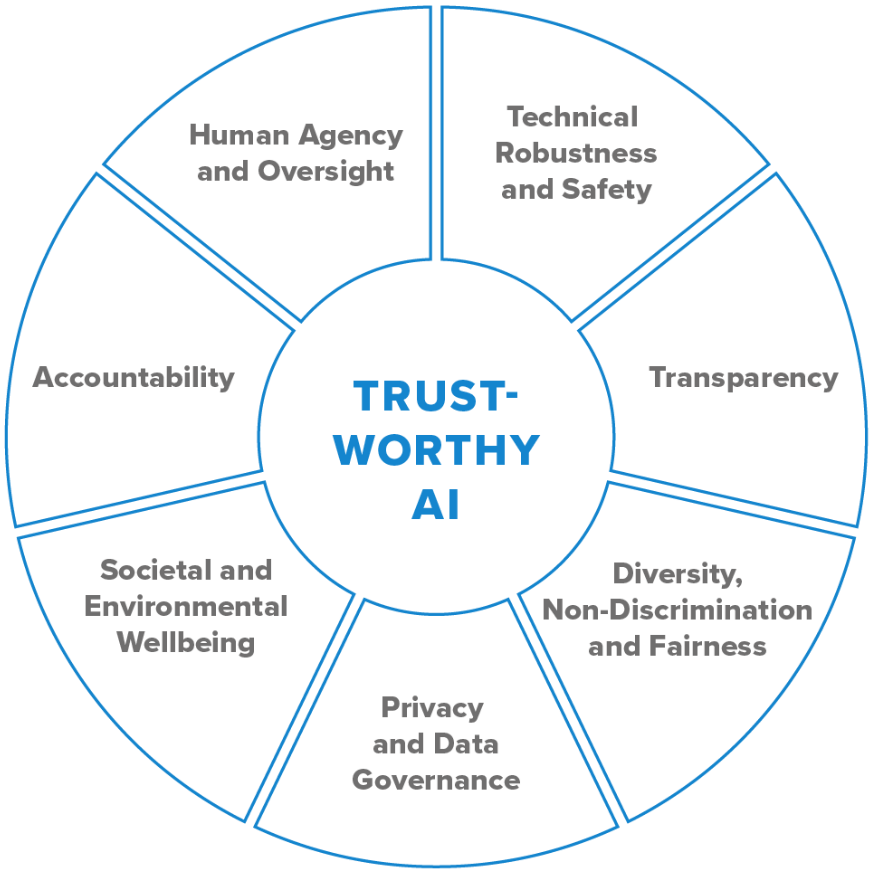
After 24 months
(mid 2026)
Requirements for High Risk Systems apply

After 36 months
(mid 2027)
Requirements for all risk systems apply

*“AI is a* ***machine-based system*** *designed to operate with* ***varying levels of autonomy*** *and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives,* ***infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions*** *that can influence physical or virtual environments.”* → very broad and includes logistic regression up to deep learning

# Trustworthy AI Dimensions

## Dimensions According to EC (2021)




https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

# Trustworthy AI Dimensions

## Dimensions According to EC (2021)

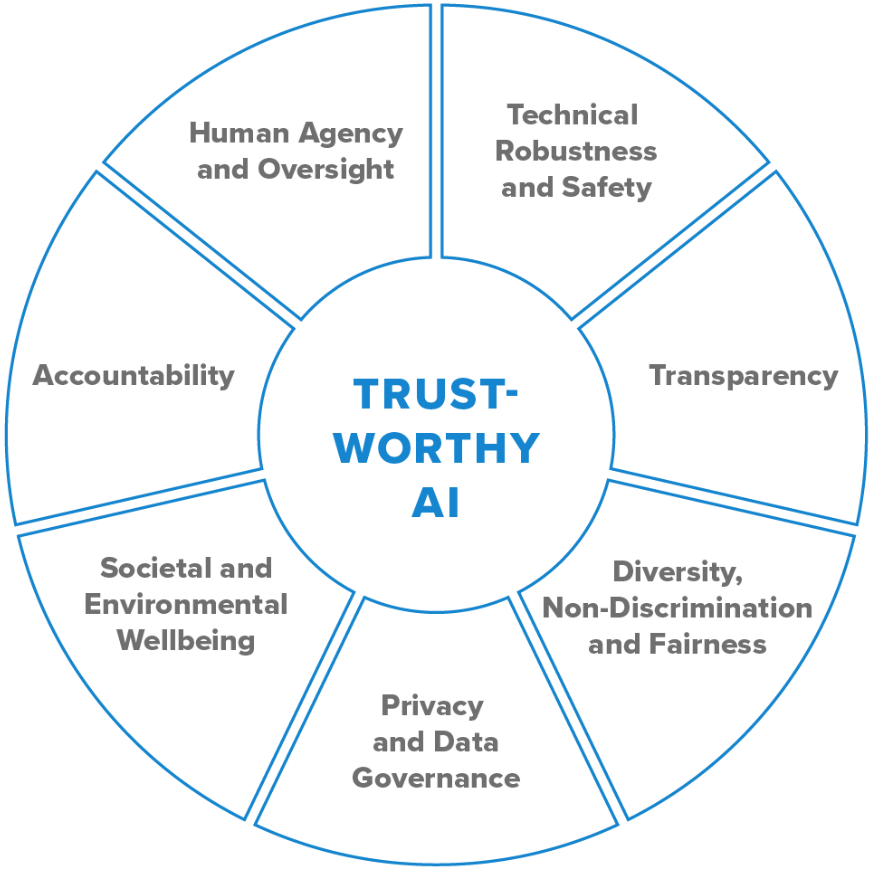
**Similar definitions and categorizations, e.g., in:**

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, *55*(2), 1-38.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys*, *55*(9), 1-46.

Human-centric & legal aspects

**Processes & guidelines**




Technical & reliability aspects

**Metrics & services**

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

An important prerequisite to validate the trustworthiness of AI, is the reproducibility of AI/ML

# Trustworthy AI Dimensions

## Dimensions According to EC (2021)

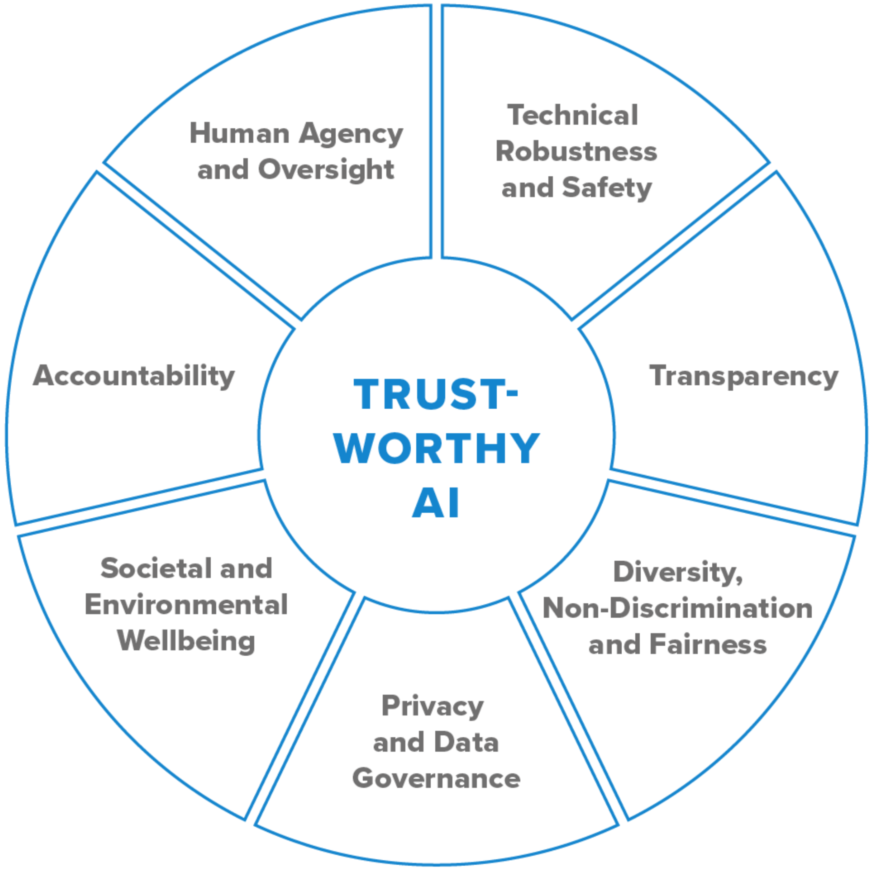
**Similar definitions and categorizations, e.g., in:**

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, *55*(2), 1-38.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys*, *55*(9), 1-46.

Human-centric & legal aspects

**Processes & guidelines**




Technical & reliability aspects

**Metrics & services**

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

# Reproducibility in AI and ML-driven Research

Semmelrock, H., Kopeinik, S., Theiler, D., Ross-Hellauer, T., & Kowald, D. (2023). Reproducibility in Machine Learning-Driven Research. *arXiv preprint arXiv:2307.10320*.

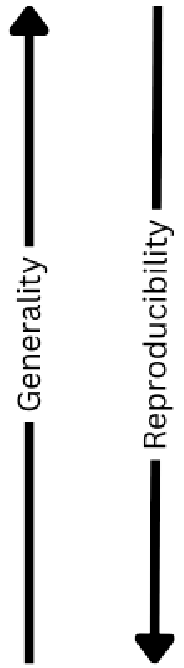
Focus on scientific fields of Computer Science and Health / Life Science

Updated version by end of May

Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A*, *379*(2197), 20200210.

# Definition of AI/ML Reproducibility

**According to Gundersen (2021)**




| | Text | Code | Data |
|---|---|---|---|
| R1 Description | | | |
| R2 Code | | | |
| R3 Data | | | |
| R4 Experiment | | | |

# AI/ML Reproducibility vs. Replicability

According to Association of Computing Machinery (ACM)

## Reproducibility

- **The results can be obtained by a different team with the same experimental setup**

## Replicability

- **The results can be obtained by a different team with a different experimental setup**

https://www.acm.org/publications/policies/artifact-review-badging

# AI/ML Reproducibility vs. Replicability

According to Association of Computing Machinery (ACM)

## Reproducibility

- **The results can be obtained by a different team with the same experimental setup**
- Refers to R4 (Experiment)

## Replicability

- **The results can be obtained by a different team with a different experimental setup**
- Refers to R1 (Description)

- R2 and R3 in between

https://www.acm.org/publications/policies/artifact-review-badging

# What are the barriers?

**R1 Description**

## 1. Completeness and quality of reporting

- Training procedure of ML model is not documented
- Evaluation metrics are not properly specified
- Evaluation results are selectively reported (e.g., for the best test-run)

## 2. Spin practices

- Inconsistency of study results and conclusions
- e.g., baseline models are used that do not fit the task → makes the own method appear stronger → findings are not reproducible

# What are the barriers?

**R2 Code**

## 3. Limited access to code

- < 1/3 of ML/AI papers share their code
- No time to polish code → do not want that others see my code
- Often only code of own model is shared but no code for baselines, evaluation metrics, etc

→ Complete pipeline needs to be shared!

# What are the barriers?

**R3 Data**

## 4. Limited access to data

- Privacy reasons (e.g., industrial setting, or sensitive data like health)
- Sensitive data can be inferred even if data is anonymized
- Often the data is shared but not the train / validation / test split

## 5. Data leakage

- Over-optimistic results due to methodological errors in train / test splits (use of test data in training process)
- Train/test split is done correctly but temporal leakage is given: timestamps in train set > timestamps in test set
- Train/test set is not independent, e.g., same person is in train AND in test set

## 6. Bias

- Biased ML models do not generalize well → issue for reproducibility
- E.g., selection bias: use data that is not representative for research question
- e.g., create test set in a specific way that favors your model

# What are the barriers?

**R4 Experiment**

## 7. Inherent non-determinism

- Often ML model outputs differ between test runs
- Sources of randomness in training process / random subsampling in k-fold cross validation

## 8. Environmental differences

- Different GPUs or CPUs lead to different results
- Different compiler versions or software versions (e.g., Java 8 vs. Java 9)

## 9. Limited access to computational resources

- Datasets too large to be calculatable on local machine → expensive server needed
- Transformers / large language models → billions of parameters to be optimized
- Reproduction costs could go to 1 - 3 Million USD

Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. arXiv website. https://arxiv.org/abs/1906.02243

KNOW Center

# What are the drivers?

**Technology-driven**

## 1. Hosted services

- Services with given runtime environment, in which models/experiments can be provided
- Limit on dataset size and computational resources
- Still different services could give different results
- E.g., Google Collab

## 2. Virtualization

- Virtual environment that can easily be shared
- Ensures that same software versions are used
- E.g., Docker images

## 3. Managing sources of randomness

- Use of fixed random seeds
- Also randomizations on hardware levels, e.g., in GPUs for parallel computations

# What are the drivers?

## Technology-driven

## 4. Privacy-preserving technologies

- Model can make accurate predictions without using the actual privacy-sensitive data
- E.g., differential privacy adds noise to the data → accuracy/privacy trade-off

## 5. Tools and platforms

- Use of frameworks such as scikit-learn instead of implementing models from scratch
- Tools like Ml-flow provide support for sound model evaluations and comparisions
- Github to share source code
- Zenodo to share data artifacts

# What are the drivers?
**Procedural**

## 6. Standardized datasets and evaluation

- Provision of benchmark datasets with defined train / validation / test splits
- Libraries with standard implementations of evaluation metrics, e.g., RecBole for recommender systems
- For novel problems (e.g., large language models), no benchmark datasets are available

## 7. Guidelines and checklists

- FAIR data principles (findable, accessible, interoperable, reusable)
- ML reproducibility checklists by conferences or journals (e.g., data shared y/n, code shared y/n)

## 8. Model info sheets

- More detailed and technically-sound version of checklists for ML models
- Also address issues like data leakage

# What are the drivers?
## Awareness

## 9. Publication policies and initiatives

- Having reproducibilty as one of the major points for paper reviewing

→ if it is not reproducible, it will not be accepted

- Pre-registration (register methodology before doing the study) becomes more relevant in ML/AI (e.g., Transactions on Recommender Systems journal)
- Initatives such as PapersWithCode.com to increase visibility of reproducible ML/AI research
- Reproducibility tracks at conferences (e.g., European Conference on Information Retrieval) to foster the reproduction of papers

# Drivers-Barriers Mapping

| Barriers \ Drivers | | Technnology-driven | | | | | Procedural | | | Awareness |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Hosted services | Virtualization | Managing sources of randomness | Privacy-preserving technologies | Tools and platforms | Standardized datasets and evaluation | Guidelines and checklists | Model info sheets | Publication policies and initiatives |
| R1 Description | Completeness and quality of reporting | | | | | | | X | | |
| | Spin practices | | | | | | | | | X |
| R2 Code | Limited access to code | X | X | | | X | | | | |
| R3 Data | Limited access to data | | | | X | X | X | X | | |
| | Data leakage | | | | | | X | | X | |
| | Bias | | | | | | X | X | | |
| R4 Experiment | Inherent nondeterminism | | | X | | | | | | |
| | Environmental differences | X | X | | | | | | | |
| | Limited access to computational resources | X | | | | | | | | |

# Conclusion and some suggestions

## Conclusion

- **Reproducibility is a prerequisite to validate the trustworthiness of AI**
- Four levels of reproducibility in AI/ML-driven research
- Barriers and drivers … but they can be mapped

## Suggestions (for the start)

- **Share your source-code via Github** (in case of double-blind review → anonymous Git repo.)
- **Share your datasets via Zenodo** (in case of double-blind review → blank authors and fill afterwards)
- **Reproducibility tracks are a great way to deal with reproducibility and to get started on doing research in a new field** (e.g., by reproducing one of the most important papers and expanding it to new domains)

Kowald, D., Schedl, M., & Lex, E. (2020). The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *Proceedings of the 42nd European Conference on Information Retrieval (ECIR'2020)*. Springer.

# Thank you!

**QUESTIONS / COMMENTS?**

**CONTACT:**

DR. DOMINIK KOWALD

DOMINIK.KOWALD.INFO

DKOWALD@KNOW-CENTER.AT

KNOW-CENTER GMBH
Research Center for Data-Driven Business and Big Data Analytics

Sandgasse 36/4, 8010 Graz, Austria
Firmenbuchgericht Graz
FN 199 685 f
UID: ATU 50367703
www.know-center.at