Challenges in pursuing AI Fairness

Scientific conference of the Greek Data Protection Authority:
1st day of Dialogue with the Research Community
Wednesday, October 1

Giorgos Giannopoulos
Maria Psalla
Lukas Kavouras
Dimitris Sacharidis
Ioannis Emiris

ATHENA Research & Innovation Information Technologies

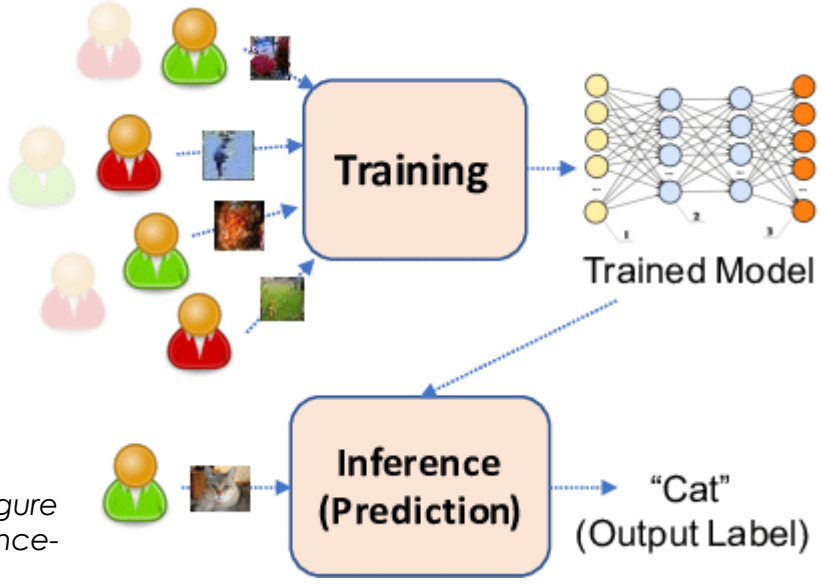
HUMANCOMPATIBLE.ORG
HUMAN-COMPATIBLE AI WITH GUARANTEES

Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or DG Connect. Neither the European Union nor DG Connect can be held responsible for them.

How can AI/ML be unfair?

How Machine Learning works?

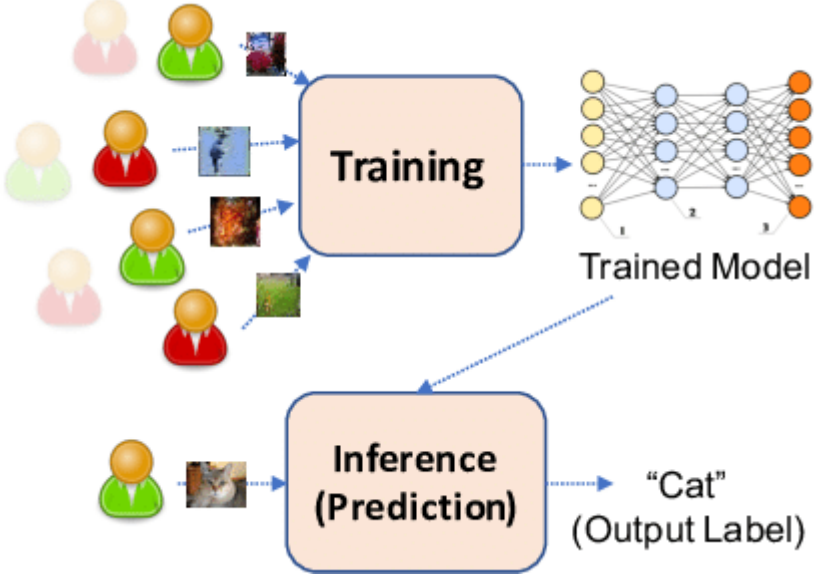
- Historical, labeled data are used to train a ML model
  - Data: Instances/records/items/individuals
  - Accompanied with a label
- The trained model is deployed on new, unlabeled data to provide labels



https://www.researchgate.net/figure/Overview-of-training-and-inference-in-deep-learning_fig1_330842645

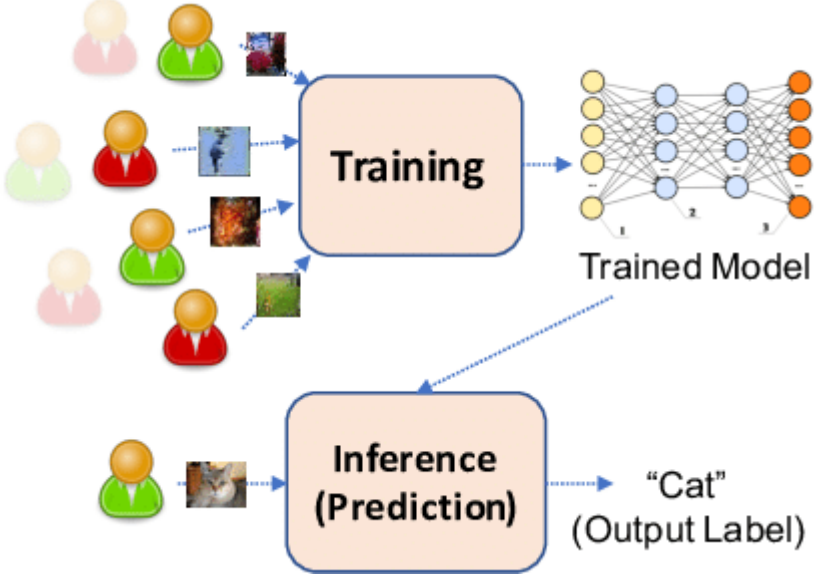
How Machine Learning works?

- A label might be:
  - The type of entity (cat, dog, human) that is depicted in an image
  - Whether you are hired/granted a loan or not
  - How much rain we expect tomorrow in Brussels
  - A recommendation score for an item for a specific person
  - …



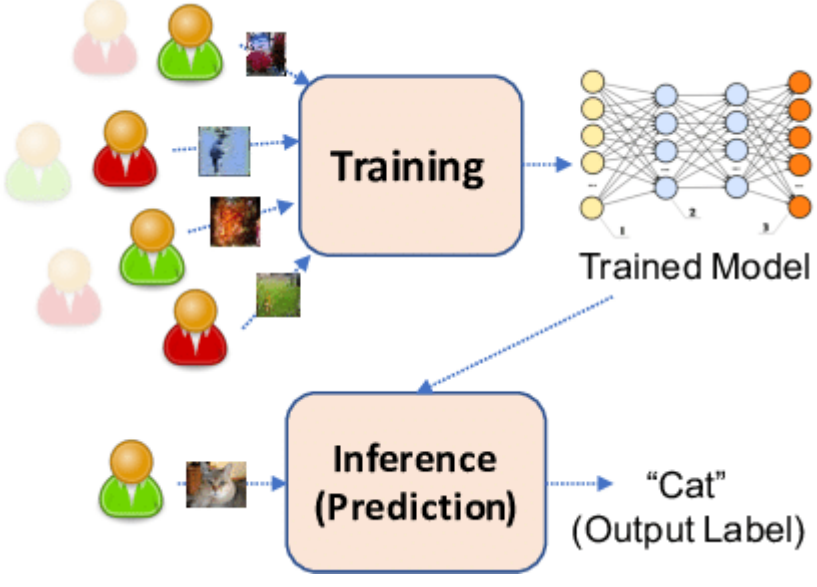
How Machine Learning works?

- A label might be:
  - Whether you are **hired/granted a loan or not**
  - Let's focus in the binary classification setting:
    - 0 or 1
    - Reject or accept
    - Unfavorable or favorable outcome



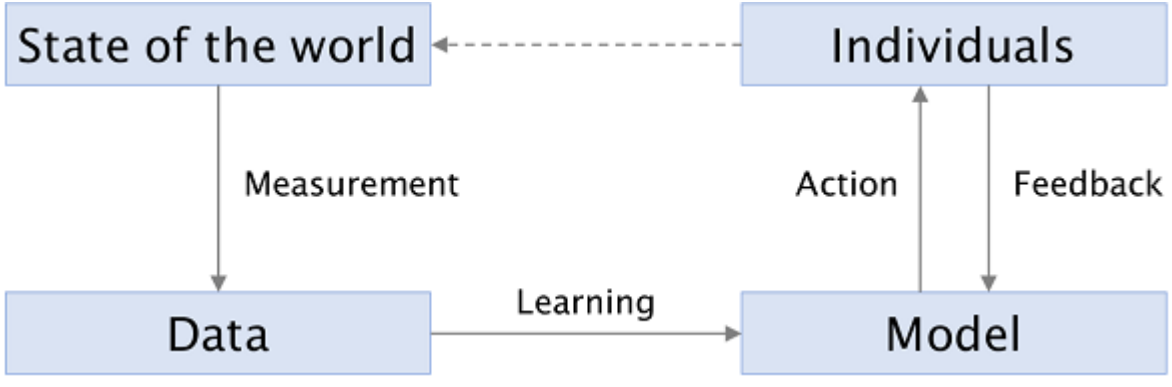
How Machine Learning works?

- A ML model tries to identify patterns:
  - Relating characteristics (features) of instances with labels



The machine learning loop

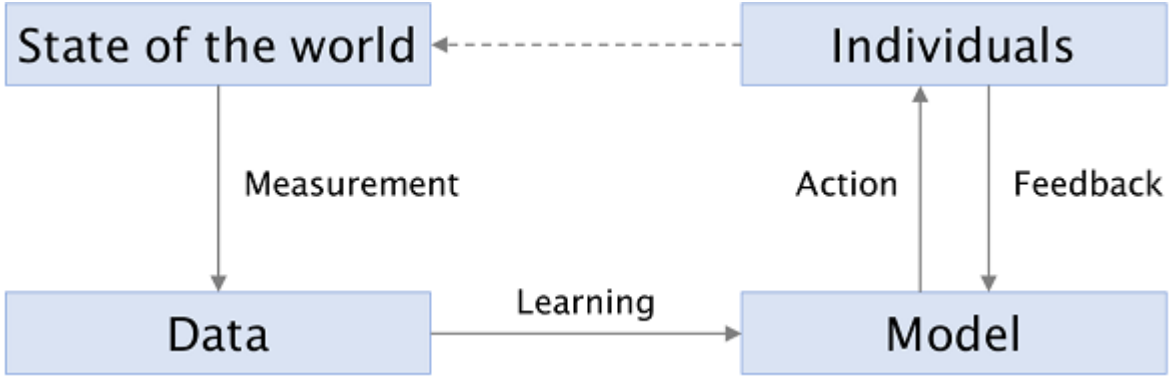
- What do we expect from a good machine learning (ML) model?



Solon Barocas et al. Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023.

The machine learning loop

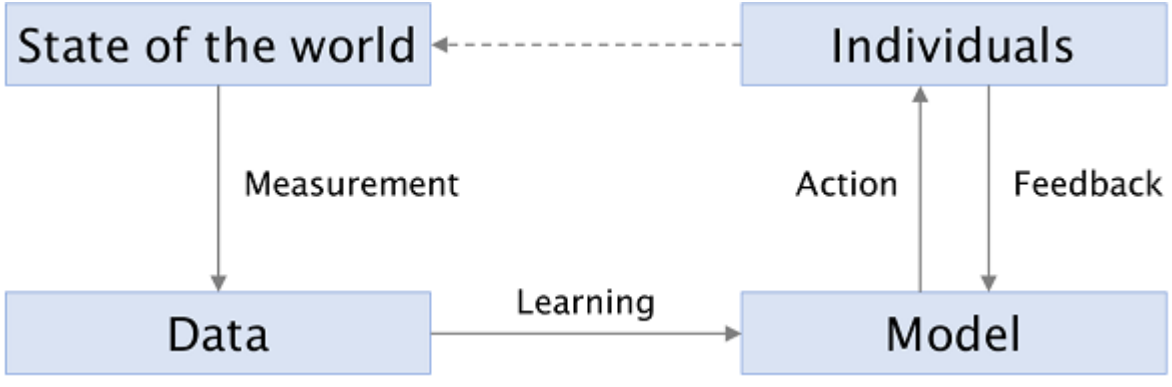
- What do we expect from a good machine learning (ML) model?
- Answer: to correctly learn (a part of) the world



The machine learning loop

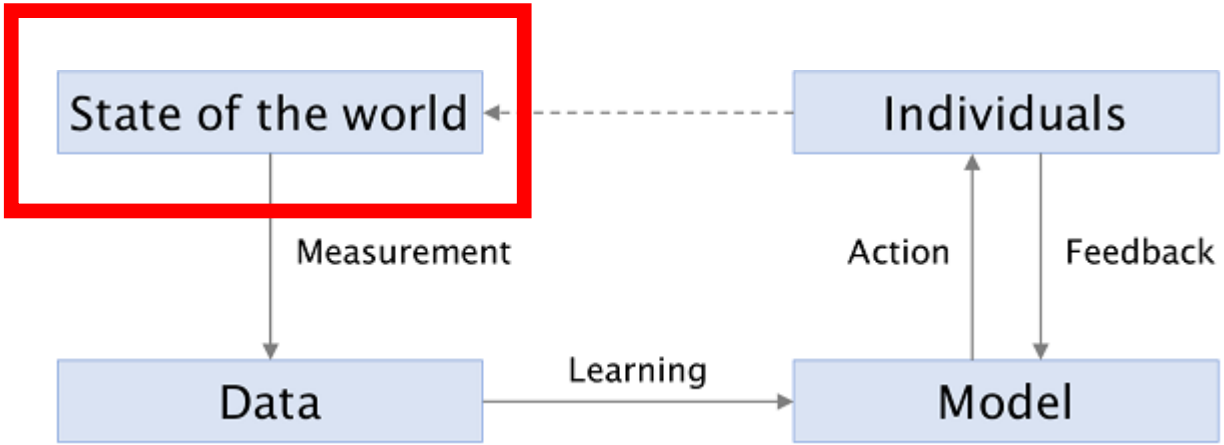
- What do we expect from a good machine learning (ML) model?
- Answer: to correctly learn (a part of) the world

- But how correct is the world?



The machine learning loop

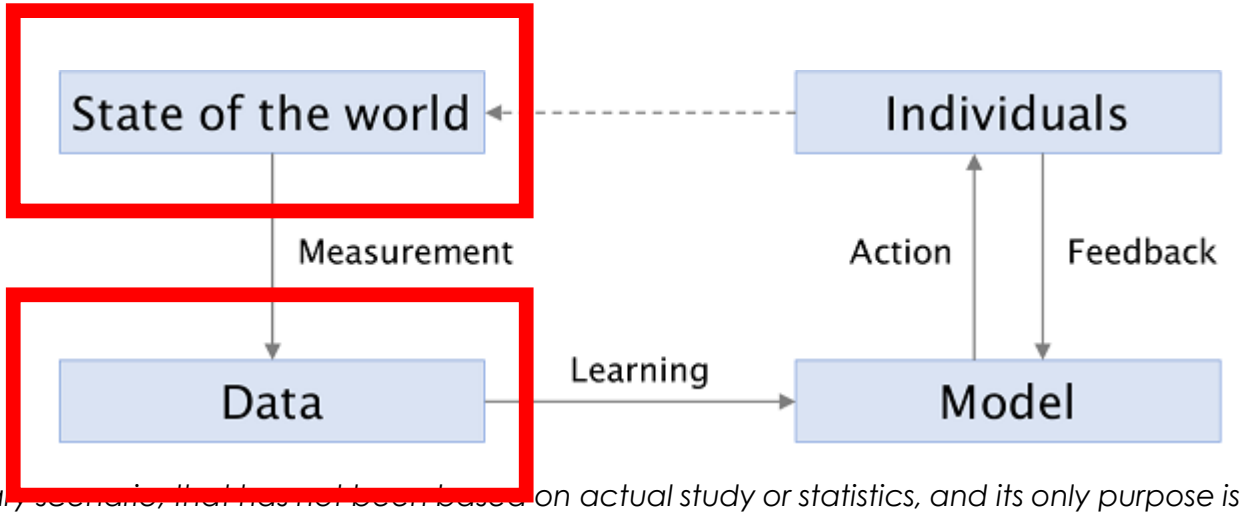
- How can we imaging the state of the "hiring in ICT companies" in, let's say, 1990?
- *The boss of the SME company single-handedly checks the submitted resumes*
  - *They believe that, since men are better at computers, they should invest **exclusively in hiring male candidates***
  - *They are sure that foreigners come from countries with lower educational level, so they **mostly ignore their resumes***



This is an imaginary scenario, that has not been based on actual study or statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

The machine learning loop

- How can we imaging the state of the "hiring in ICT companies" in, let's say, 1990?
- This has a twofold effect
  - It creates historical, training data where **non-males** and individuals from **other countries** consistently get the **0/reject label**
  - It **discourages the specific subpopulations** to study or further pursue a career in ICT → in 2000, less non-males will pursue ICT jobs

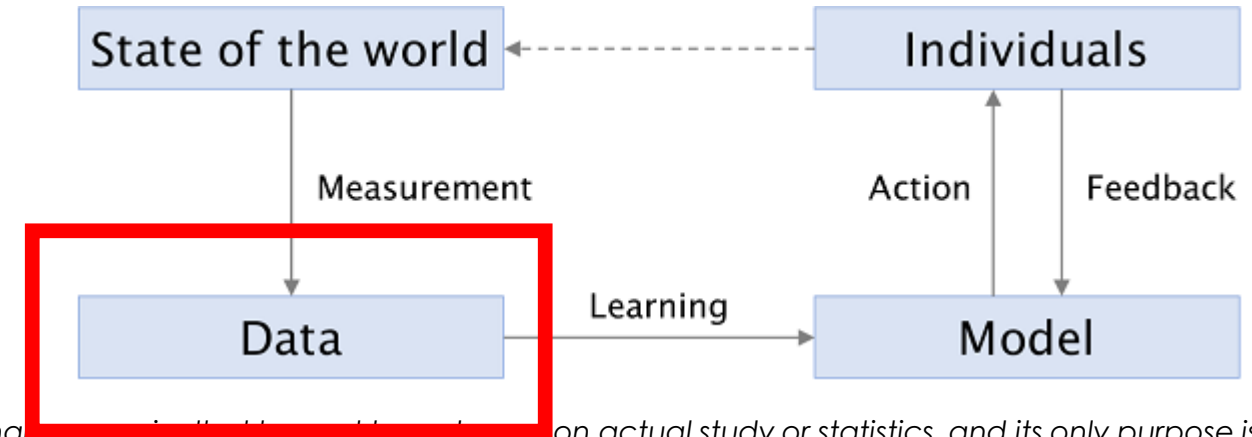


This is an imaginary scenario, that has not been based on actual study or statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

The machine learning loop

- How can we companies" in, let's say, 199
- This has a tw
  - It creates h nd individuals from other cour
  - It discoura further pursue a career in I obs

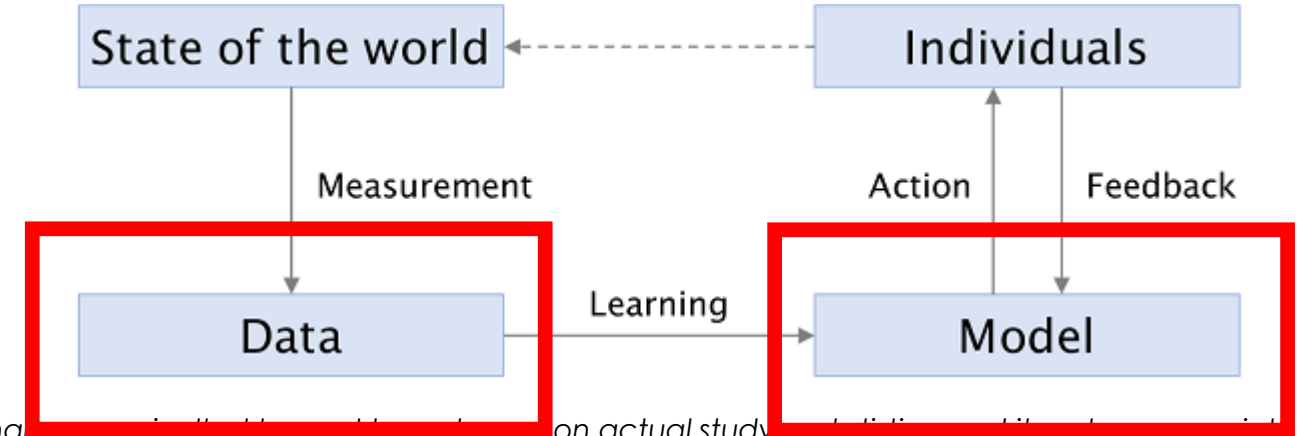
Historical/structural bias has been inserted in job hiring training data



This is an imaginary scenario, that has not been based on actual study or statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

The machine learning loop

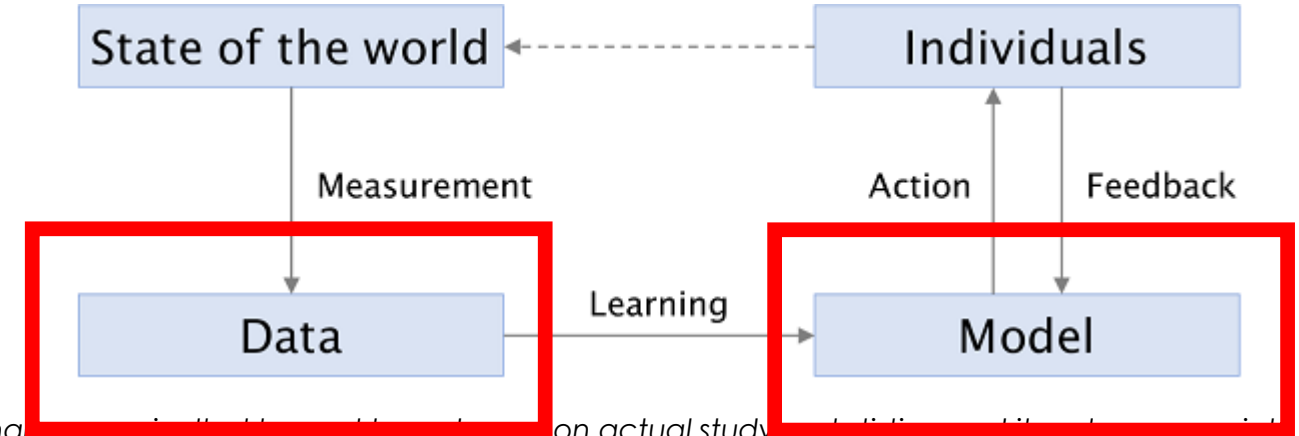
- What do we expect a good ML to learn in 2024?
- Train on historical hiring data from the last 30 years
  - Learn to relate features of candidates and jobs to reject/accept labels



This is an imaginary scenario, that has not been based on actual study of statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

The machine learning loop

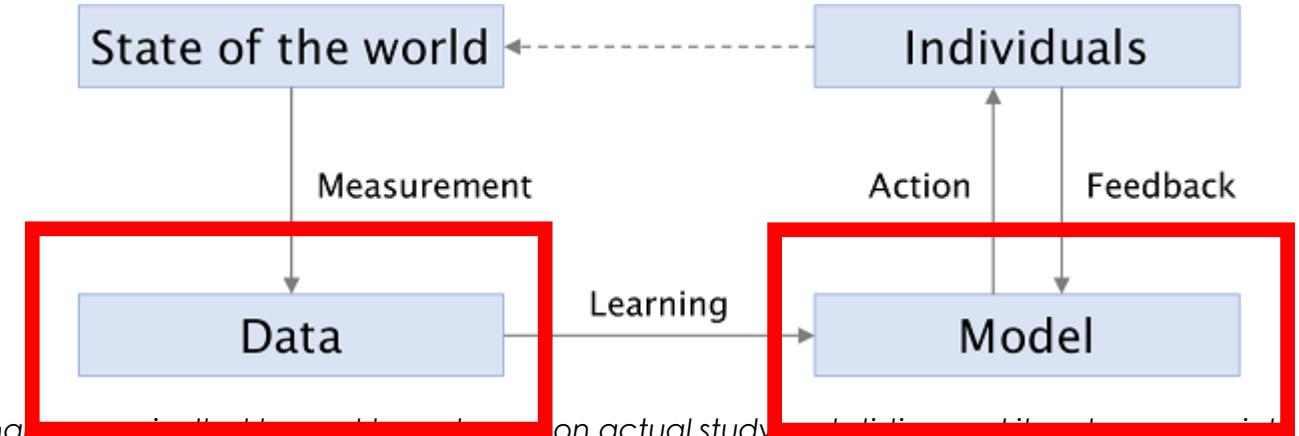
- *Learn to relate features of candidates and jobs to reject/accept labels*
  - <UniA, 3yearsexperience, C++, male> → accept
  - <UniB, 3yearsexperience, Java, female> → reject
  - …



This is an imaginary scenario, that has not been based on actual study of statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

The machine learning loop

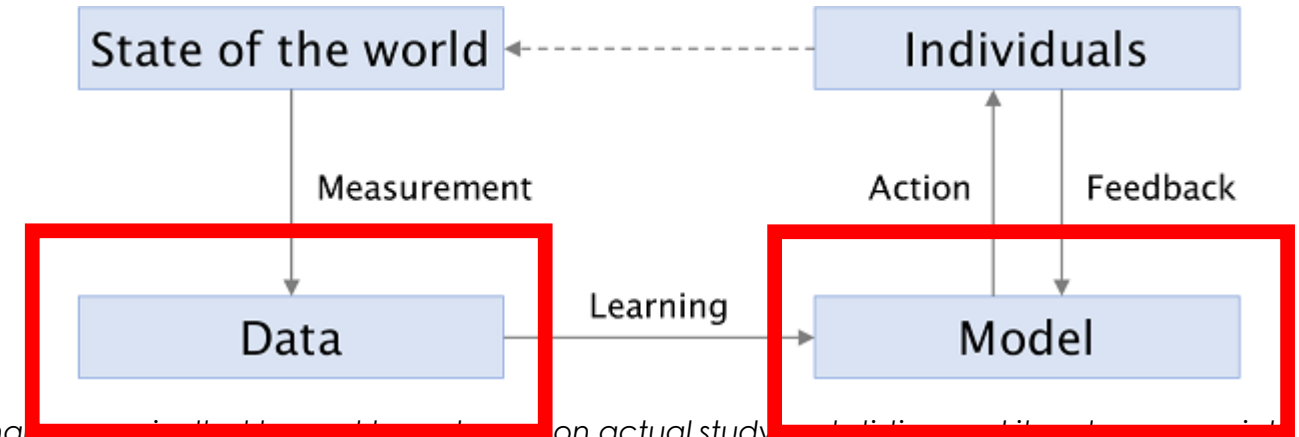
- *Learn to relate features of candidates and jobs to reject/accept labels*
  - <**UniA**, 3yearsexperience, C++, **male**> → accept
  - <**UniB**, 3yearsexperience, Java, **female**> → reject
  - …



This is an imaginary scenario, that has not been based on actual study of statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

The machine learning loop

- *Learn to relate features of candidates and jobs to reject/accept labels*
  - <**UniA**, 3yearsexperience, C++, **male**> → **accept**
  - <**UniB**, 3yearsexperience, Java, **female**> → **reject**
  - …

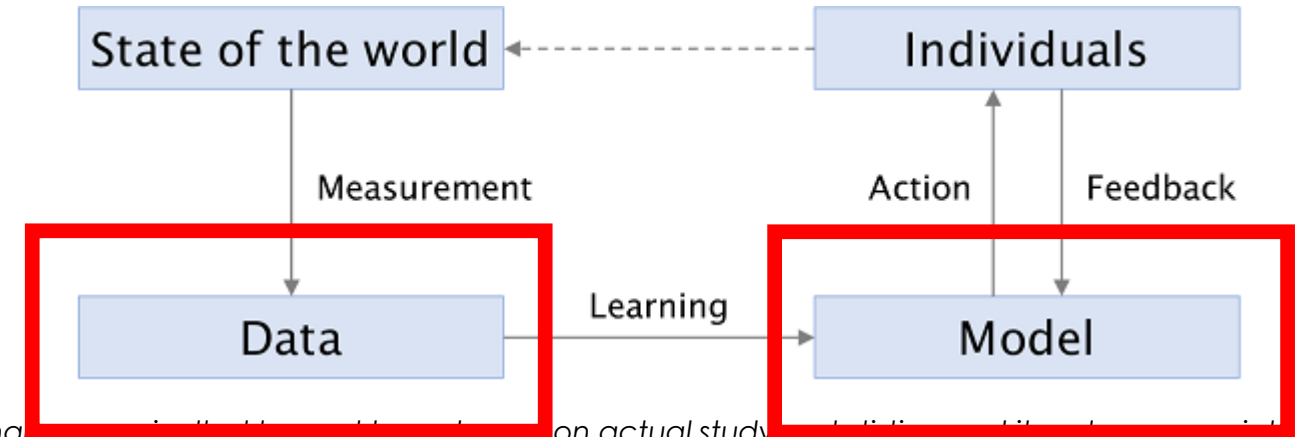


This is an imaginary scenario, that has not been based on actual study of statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

The machine learning loop

- *Learn to rel… …o reject/accept labels*
  - <**UniA**, 3ye…
  - <**UniB**, 3ye…
  - …

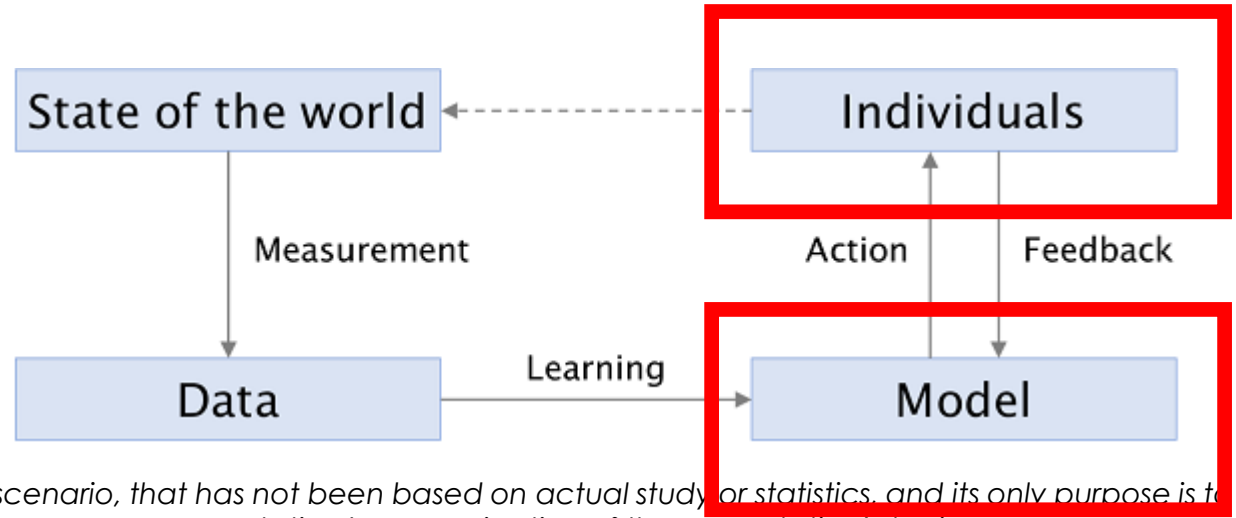
Proxy discrimination



This is an imaginary scenario, that has not been based on actual study of statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

The machine learning loop

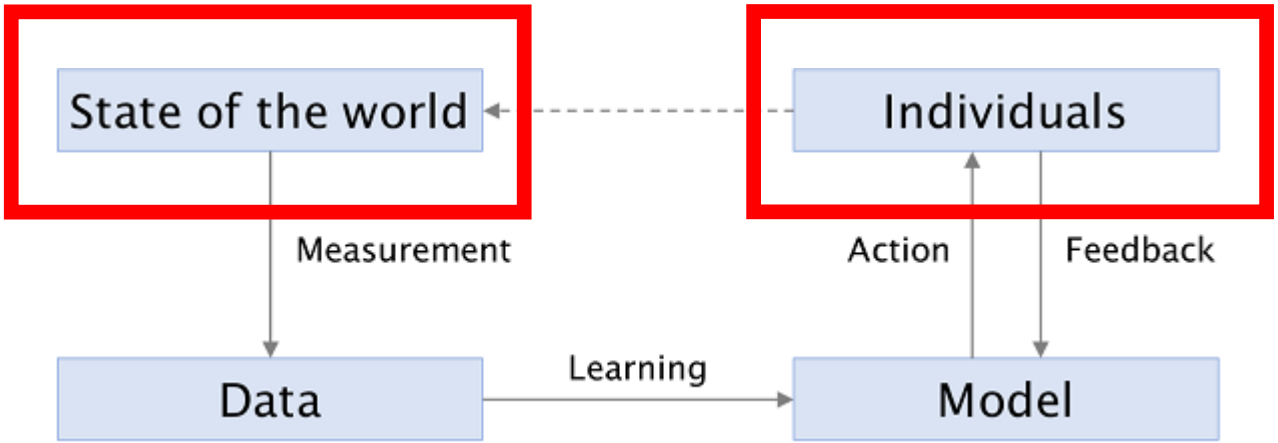
- Individuals will receive biased decisions



This is an imaginary scenario, that has not been based on actual study or statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

The machine learning loop

- Individuals will receive biased decisions
  - Sensitive groups will be discriminated against
  - Bias will be perpetuated → feedback loop

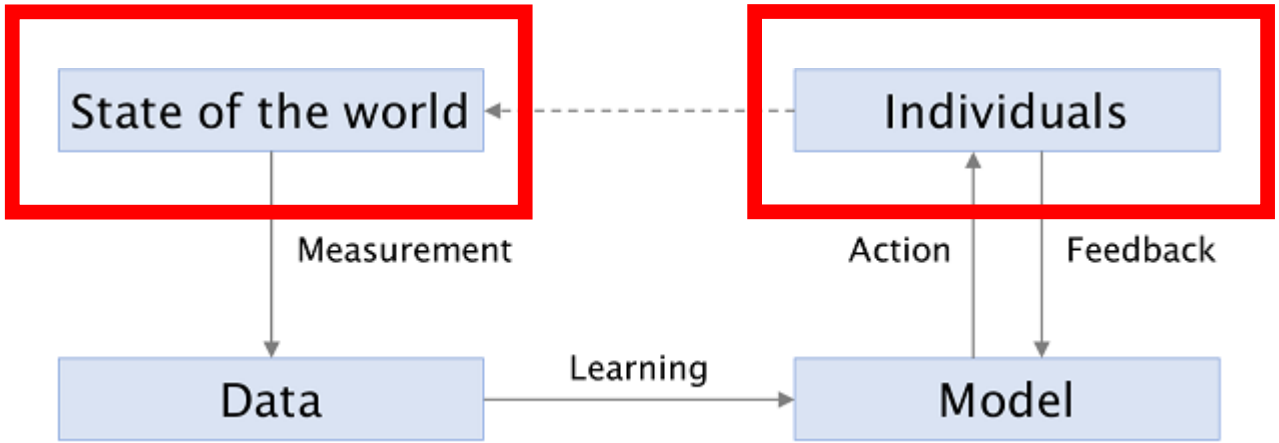


This is an imaginary scenario, that has not been based on actual study or statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

The machine learning loop

- Individuals v
  - Sensitive g
  - Bias will be

Feedback loops



This is an imaginary scenario, that has not been based on actual study or statistics, and its only purpose is to facilitate the presentation/communication of the presentation's topics

Discrimination in law

Non-discrimination law

- The aim of non-discrimination law is to allow all individuals an equal and fair prospect to access opportunities available in a society

- **Discrimination in the EU Law**
  - The law of the Council of Europe
  - The law of the European Union

  - Six protected attributes:
    - racial and ethnic origin, sex, religion and belief, disability, age and sexual orientation

- **Discrimination in the US Law**
  - The Congress (legislature)
  - The Courts (judiciary)
  - Federal Agencies (executive governance)

Gap between law and algorithms

- A significant gap between law and algorithms, consists in the **process-oriented nature** of (EU) law, versus the **result-oriented nature** of algorithmic systems.
- In a **process-oriented assessment**, each individual case would be examined in detail
  - e.g. "a judge would need to evaluate whether the formulated requirements are appropriate and the resulting bias is acceptable in this specific case".
- This is **not feasible in an algorithmic setting**, since an AI system has learnt to **model patterns over large datasets**, i.e. large number of individual cases,
  - a result of an AI system cannot adequately represent all the specifics and the context of the respective case under examination.

- Legal perspective on possible fairness measures –A legal discussion using the example of hiring decisions - Hauer et al. 2021

Gap between law and algorithms

- "To date, there has been **no legal definition of fairness beyond individual decisions of jurisdiction**. This is mainly due to the process-oriented assessment of equality.
- With the upcoming algorithm-based decision making though, it will be essential to find a clear specification of fairness as the assessment will no longer work on a process level.
- Rather, **there will be a decisive shift to a result-based assessment**. It is therefore of utmost importance to analyze, whether and how fairness measures comply with EU anti-discrimination legislation"

- Legal perspective on possible fairness measures –A legal discussion using the example of hiring decisions - Hauer et al. 2021

Bridging the legal and the algorithmic view

Challenges and criteria

- Important issues and shortcomings that law had not considered and came up or where emphasized with the advent of ML based decision making
  - **Equal opportunity vs equal outcome**
  - **Proxy variables and correlations**
  - **Intersectional-subgroup fairness**
  - **Feedback loops**
  - **Robustness**
  - Sampling requirements
  - Computational complexity
  - Explainability

Equal opportunity vs Equal outcome

- **Equality of opportunity** prescribes that all individuals are given the same chances to achieve a favorable outcome.
  - For example, a hiring recommendation system should assign an "accept" label to candidates based on objective criteria and labeled training data, that do not take into account sex information in its decision.
- **Equality of outcome** prescribes that all protected (sub)groups equally/proportionally obtain the favorable outcome.
  - For example, the ratio of males to females that obtain the "accept" label should be proportionate to the respective candidate's ratio, even if this conflicts with the rating produced by the recommender.

Equal opportunity vs Equal outcome

- **Equality of outcome** is a notion that is based on the recognition of structural inequalities and historical bias in procedures and datasets
  - It is pursued via instruments such as **affirmative** action (or positive action, positive discrimination),
  - Aims at alleviating/eliminating such inequalities against sensitive subpopulations.
  - For example, affirmative action or a company's policy would require a minimum quota in female "acceptances" for every job type

Equal opportunity vs Equal outcome

- Deciding on the trade-off between **equal treatment** (or equality of opportunity) versus **equal outcome** is important
- The authors map this trade-off to a dilemma between **formal** equality and **substantive** equality
  - **Formal** equality aims at ensuring that a system does not introduce additional bias to each decisions
  - It includes the implicit assumption that the status quo is fair, i.e. there does not exist historical/structural bias in the data
  - **Substantive** equality accounts for historical biases and aims at reducing them
- The authors claim that the goal of the EU non-discrimination law is actually *substantive equality*

Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non Discrimination Law - Wachter et al 2021

Equal treatment vs Equal outcome

- Categorization of existing algorithmic fairness definitions into bias **preserving** and bias **transforming**
  - Statistical measures that take into account **solely the model's predictions** in measuring fairness are categorized as **bias transforming**
  - Measures that take into account **both predictions and actual labels** (existing in historical, labeled datasets) are considered as **bias preserving**
  - Measures that examine or account for, **causal relations between attributes and outcomes** including counterfactual-based methods, are also categorized as **bias transforming**

Fairness metric	Bias preserving?
1. Group fairness, Statistical (demographic) parity	✗
2. Conditional statistical (demographic) parity, Conditional independence	✗
3. Predictive parity, outcome test	✓
4. False positive error rate balance	✓
5. False negative error rate balance, Equal opportunity	✓
6. Equalized odds	✓
7. Conditional use accuracy equality	✓
8. Overall accuracy equality	✓
9. Treatment equality	✓
10. Test-fairness or calibration	✓
11. Well-calibration	✓
12. Balance for positive class	✓
13. Balance for negative class	✓
14. Causal discrimination (direct discrimination)	*
15. Fairness through unawareness	*
16. Fairness through awareness	✗
17. Counterfactual fairness	✗
18. No unresolved discrimination	✗
19. No proxy discrimination	✗
20. Path based causal reasoning	✗

Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non Discrimination Law - Wachter et al 2021

Proxy variables and correlations

- **Proxy discrimination**: bias is expressed not directly via sensitive attributes, but indirectly, via proxy variables, that are to some extent correlated with the respective sensitive attribute
- Examples
  - height and maternity leave attributes serving as proxies for the sex sensitive attribute
  - university serving as proxy for the sex sensitive attribute
  - residence/location attributes serving as proxies for the race sensitive attribute, etc.

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. California Law Review, 104:671, 2016

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In KDD '08.

Proxy variables and correlations

- **Fairness by unawareness**: a commonly encountered misunderstanding
  - If sensitive attributes are excluded from an AI model's training, fairness is ensured

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. California Law Review, 104:671, 2016

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In KDD '08.

Proxy variables and correlations

- **Fairness by unawareness**: a commonly encountered misunderstanding
  - If sensitive attributes are ex[illegible]rom an AI model's training, fairness is ensured
  - bias can be perpetuated via proxy discrimination. That is, even if sensitive attributes are removed, the bias of the training data can still be transferred into the trained model

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. California Law Review, 104:671, 2016

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In KDD '08.

Intersectional-subgroup fairness

- **Intersectional bias** arises when considering subpopulations defined by more than one attribute, with at least one of them being a sensitive one.
- Challenges
  - it is often the case that **bias is magnified** for specific subgroups;
    - due to **preexisting discrimination** towards individuals belonging to these groups
    - due to their **under-representation** in datasets
  - data sparsity → **uncertainty in evaluating bias** for these subgroups when auditing a dataset or algorithm:
    - since very few instances representing a specific subgroup might be found in an audited dataset, the **significance of the findings can be questionable**
  - computational issues arise when trying to drill down to more granular subgroups, since **complexity increases exponentially**

Arjun Roy, Jan Horstmann, and Eirini Ntoutsi. Multi-dimensional discrimination in law and machine learning - a comparative overview. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23.

Algorithmic discrimination in Europe Challenges and opportunities for gender equality and non-discrimination law - Gerards & Xenidis 2021

Feedback loops

- **Feedback loops** comprise self-repeating processes that can potentially reinforce and perpetuate preexisting bias
  - For example, if a hiring recommendation system is initially trained on a biased dataset, then **its recommendations will probably reproduce the bias** (if no fairness-correcting action is taken)
  - Then, these **new recommendations can be used as additional training data**, that also carry bias
  - Further, applying the system in real-world domains and continuously rejecting female candidates in favor of male ones, might **discourage individuals from the formerly protected groups from applying for specific job positions**. It is well recognized in the literature that the pattern recognition and learning mechanisms applied by many AI systems can facilitate the creation of feedback loops

Pauline T. Kim. Data-driven discrimination at work. William and Mary law review, 58:857, 2017.

Algorithmic discrimination in Europe Challenges and opportunities for gender equality and non-discrimination law - Gerards & Xenidis 2021

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. California Law Review, 104:671, 2016.

Robustness to manipulation

- Discrimination can often be **intentional**, meaning that the system/data/application owner is aware of preexisting bias and
  - does not apply any bias-correction actions or
  - even tries to hide it or
  - explicitly introduces bias
- Masking bias can be achieved through various ways,
  - E.g. **gerrymandering** or
  - **manipulating the output of fairness auditing/explainability methods** to render the audited model seemingly fair, while it is not
- The cited work of prominently demonstrates how a classifier can be retrained in an adversarial way, to
  - maintain the same level of accuracy,
  - and at the same time **suppress the explicit contribution of sensitive attributes**, so that a large set of **explainability methods are tricked** into deciding that its outputs are fair, while they are not.

B. Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In SafeAI@AAAI, 2020.

Additional

- Sampling requirements
- Computational complexity
- Explainability

Key directions

Key considerations

- **Discrimination by proxy variables, intersectional fairness and feedback loops** comprise major issues when pursuing fairness in real world applications
  - Ongoing/existing algorithmic work but no one-size-fits-all solutions exist yet
  - The current EU legal framework needs to be adapted and extended, since it is widely recognized that it has not been able to appropriately handle such issues up until now
- **No one-size-fits-all fairness definitions** or bias detection methods exist.
  - Fairness is highly **application-, scenario- and context-specific**, since different real world applications of AI decision making systems and different social circumstances highly affect what is considered as "fair"
  - Since the law cannot specialize on a case by case basis, this needs to be done by **domain experts in collaboration with governmental and independent supervising and auditing authorities**.

Key considerations

- **Cross-sectorial collaboration** is a necessity in practically every step of building both fair-by-design systems and methodologies, and AI fairness policies.
  - There exists a large gap between law/ethics and data/algorithms and only such collaboration can bridge this gap and produce meaningful policies and best practices
- **Specific fairness definitions have** been distinguished by prominent studies on the intersection of law and algorithms
  - Conditional Independence, Separation (e.g. equal opportunity, equalized odds), Sufficiency (e.g. Calibration) can be considered suitable in different the application settings and contexts,
  - **Counterfactual Fairness** is considered a sufficiently expressive and adaptable definition that allows it to generalize in different cases and optimally represent substantial equality, in the spirit of the EU law

The AutoFair project

- Horizon Europe R&D project
- Research on novel methods for detecting, quantifying, assessing, explaining and correcting bias (unfairness) in AI
  - https://humancompatible.org/index.php/about/
- Our focus: **Fairness aware explainability**

Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or DG Connect. Neither the European Union nor DG Connect can be held responsible for them.

The AutoFair project

- Our results
  - ***Fairness in AI: bridging the gap between algorithms and law.*** FAIR@ICDE 2024
  - **Fairness Aware Counterfactuals for Subgroups.** NeurIPS 2023
  - **Auditing for Spatial Fairness.** EDBT 2023
  - **FALE: Fairness-Aware ALE Plots for Auditing Bias in Subgroups.** Uncertaintly meets explainability workshop@ECML 2023

  - **https://github.com/AutoFairAthenaRC**

Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or DG Connect. Neither the European Union nor DG Connect can be held responsible for them.