visuAAL

Privacy-Aware and Acceptable Video-Based Technologies
and Services for Active and Assisted Living

# Algorithmic governance for active assisted living

## ESR11 – Sophie Noiret

PhD Seminar, Aachen

May 5th, 2011

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 861091".

Universitat d'Alacant
Universidad de Alicante
Project Coordinator

RWTH AACHEN UNIVERSITY

Stockholm University

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

TU WIEN

# About me

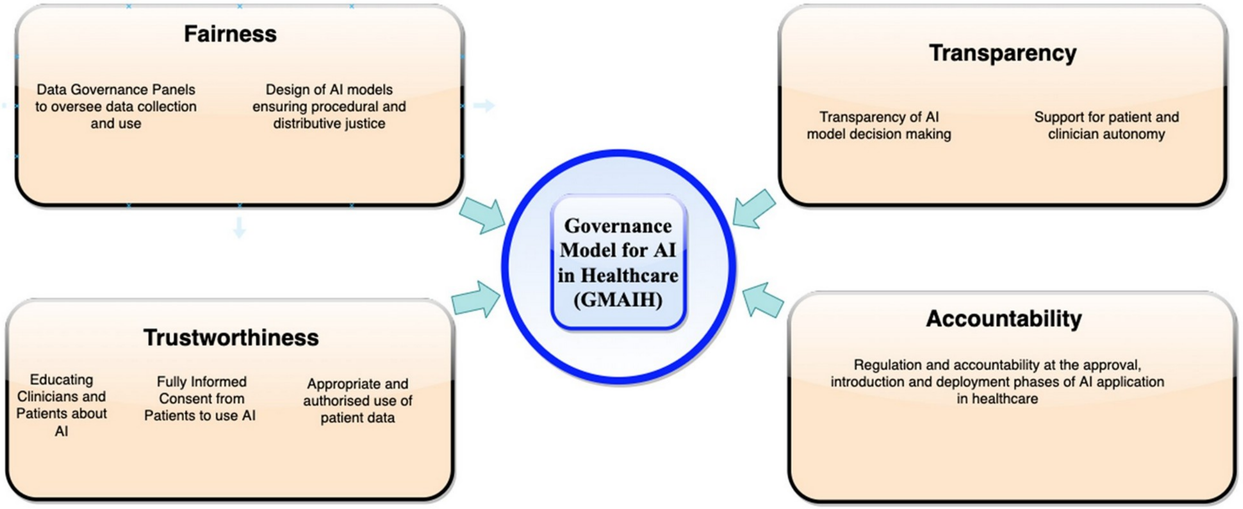
- Master's Degree in Engineering from the Ecole Centrale de Nantes (France) in 2018, with a specialty in Robotics and Embedded Systems
- Worked as a software developers in the aeronautics industry
- Started at TU Wien 1.st April 2021

# About me

- Master's Degree in Engineering from the Ecole Centrale de Nantes (France) in 2018, with a specialty in Robotics and Embedded Systems
- Worked as a software developers in the aeronautics industry
- Started at TU Wien 1.st April 2021

# Algorithmic Governance

- Machine-like in nature, and founded on computer-based procedures and rules
- From sociology and philosophy
- Technical standpoint : bias, transparency


Fairness
Data Governance Panels to oversee data collection and use
Design of AI models ensuring procedural and distributive justice
Transparency
Transparency of AI model decision making
Support for patient and clinician autonomy
Governance Model for AI in Healthcare (GMAIH)
Trustworthiness
Educating Clinicians and Patients about AI
Fully Informed Consent from Patients to use AI
Appropriate and authorised use of patient data
Accountability
Regulation and accountability at the approval, introduction and deployment phases of AI application in healthcare

S. Reddy, S.Allan, S. Coghlan, P. Cooper, A governance model for the application of AI in health care, *Journal of the American Medical Informatics Association*, Volume 27, Issue 3, March 2020, Pages 491–497

# Bias in Computer System

- Biased Computer Systems : “systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others”[1]
- Can come from:
  - Data: Not representative of reality (representation bias, measurement bias, population bias, temporal bias, etc.) ; Class imbalance
  - Algorithm: Exacerbate pre-existing imbalance
  - User Interaction:  Feedback loop, over-trusting of AI system

[1] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems.  ACM Transactions on Information Systems , Vol. 14, Issue 3 (July 1996), pp 330–347.

# Explainable AI

- Understanding and Explanation of AI systems
- Intrinsically Interpretable models: Linear Regression, Logistic regression, Decision Tree, Rule-Based, etc.
- Explanation of black-box models

# How does algorithmic unfairness manifests in and influences care ?

- Disparities in Healthcare: socioeconomic status, education status, age, gender, sexual identity/orientation, body mass index (BMI), etc.
- Known instances of algorithmic discrimination in facial analysis, recidivism predictions, hiring tools, social media
- Algorithmic bias in Healthcare: Diagnostic from X-Rays, skin lesion classification, healthcare resources allocation, etc.

# What are the factors of discrimination in care technologies ?

- Existing bias in care and medical field
- Legally protected attributes
- ML techniques for the discovery of unknown biases

## How are these factors encoded by data and algorithms used in AAL technologies ?

- Explicit presence of the attributes & proxies
- How do they influence the outcome ?

# How can this bias be identified ...?

- Opacity of algorithms: Trade secrecy, technical illiteracy, black-box nature of the algorithm
- Need for explanainability techniques
- Discovery of unknown biases

# … and quantified ?

- Definitions of fairness from legal and technical field
- Different, potentially incompatible notion of fairness
- Different metrics

# How can this bias be mitigated ?

- Pre-processing, in-processing, post-processing
- Effect of intervention at different point in the pipeline
- Trade-off between performances, fairness and explainability

# System Specific Questions

- Does the use of different modalities (RGB, depth, thermal ) plays a role ?
- Does bias in one step (e.g. person detection, joint extraction) plays a role downstrem (e.g. fall detection, activity recognition)
- Can synthetic data be used to compensate for lack of representativity in the dataset ?
- How can an explanation be given across several subsystems ?

# Assessing Algorithmic Fairness without Sensitive Information

- Fairness with Limited Awareness
  - Fair transfer learning
  - Proxy Fairness
- Fairness Without Demographics
  - Fair Class Balancing
  - Disentangled Representations
  - Optimizing for worst-case scenario

S. Noiret. 2021. Assessing Algorithmic Fairness without Sensitive Information. In Proceedings of the Conference on Information Technology for Social Good (GoodIT ’21). Association for Computing Machinery, New York, NY, USA, 325–328.

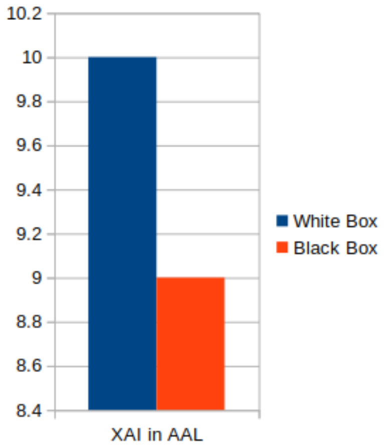
# Bias and Fairness in Computer Vision Applications of the Criminal Justice System

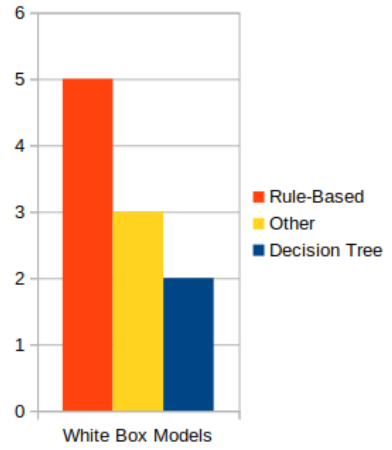
- Interview of researchers and CS
  - Description of ML strategy
  - Data
  - Explainability -  Interpretability – Fairness
- Findings:
  - Fairness was either not considered or considered « nice to have »
  - There is a gap between state-of-the art methods and real-word applications

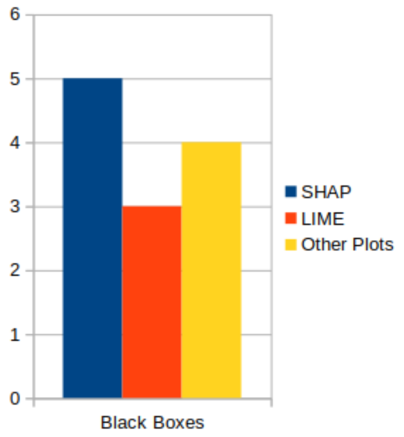
S. Noiret, J. Lumetzberger, M. Kampel. Bias and Fairness in Computer Vision Applications of the Criminal Justice System, 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 2021, pp. 1-8

# Explainable AI in AAL

- Databases: ScienceDirect, MDPI, IEEE
- Keywords:
  - XAI, Explainable AI
  - AAL, Assited Living
  - Vital signs monitoring, Lifelogging,Activity/behaviour detection/recognition, remote monitoring, gait analysis, fall detection/prevention, gesture recognition, emotional recognition, food intake

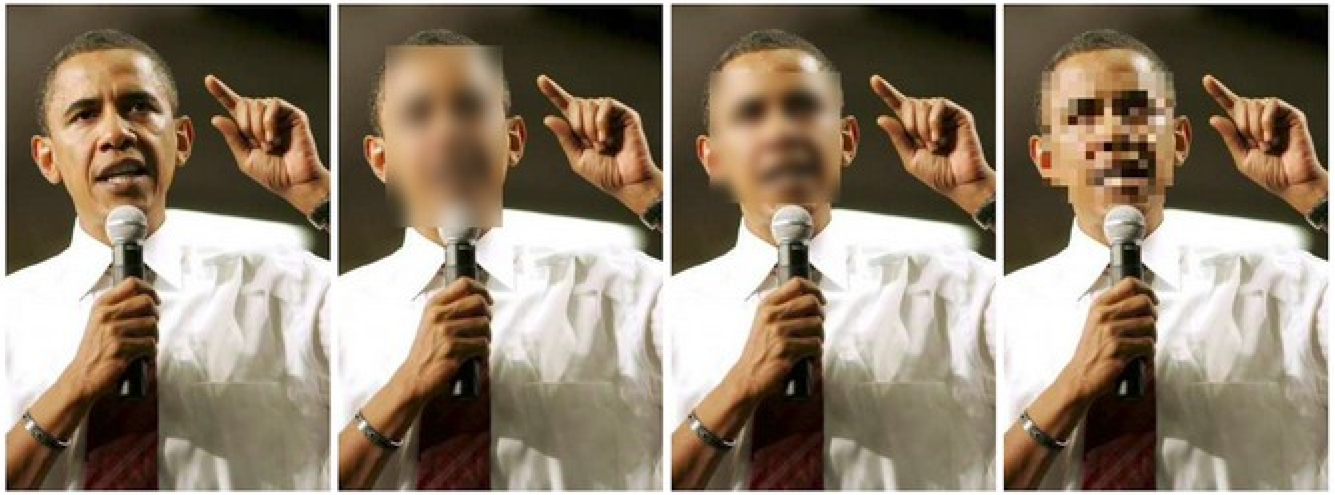






# Fairly Private: Investigating The Fairness of Visual Privacy Preservation Algorithms

- Are visual privacy preservation techniques fair?
- If PPA are biased, does it depend on the method used for obfuscation? detecting the face? recognition?

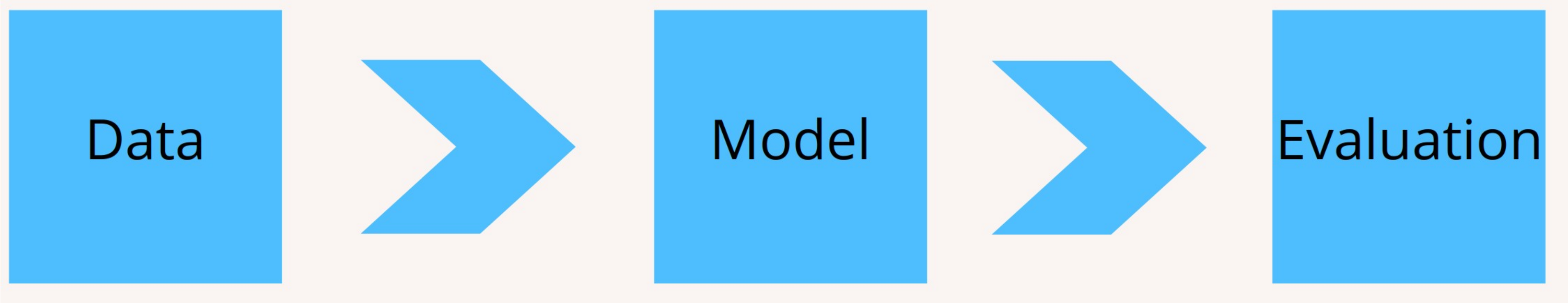


S. Noiret, S. Ravi, M. Kampel, F. Florez-Revuelta Fairly Private: Investigating The Fairness of Visual Privacy Preservation Algorithms, *submitted*

# Publications

- S. Noiret. 2021. **Assessing Algorithmic Fairness without Sensitive Information**. In *Proceedings of the Conference on Information Technology for Social Good (GoodIT ’21)*. Association for Computing Machinery, New York, NY, USA, 325–328.
- S. Noiret, J. Lumetzberger, M. Kampel. **Bias and Fairness in Computer Vision Applications of the Criminal Justice System**, *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021, pp. 1-8
- State of the Art of Audio- and Video-Based Solutions for AAL
- S. Noiret, S. Ravi, M. Kampel, F. Florez-Revuelta **Fairly Private: Investigating The Fairness of Visual Privacy Preservation Algorithms,** *acceptance pending*

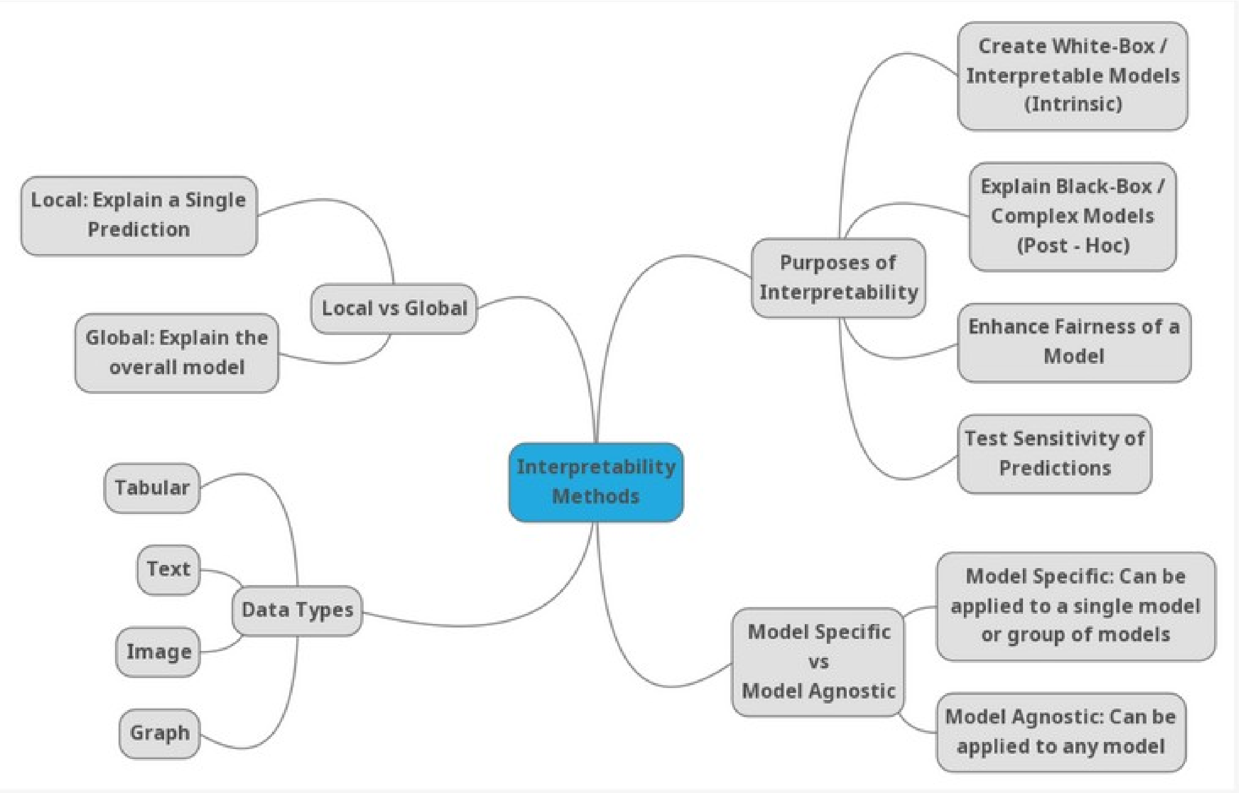
# Futur work



- Evaluation: evaluate group level fairness and create explanation for known attributes; look for unknown factors and proxies
- Intervention at data level: more  representative data; synthetic data
- Intervention at the model level: model specific and model agnostic

# Thank you!

# Explainable AI


Interpretability Methods
Local vs Global
Local: Explain a Single Prediction
Global: Explain the overall model
Data Types
Tabular
Text
Image
Graph
Purposes of Interpretability
Create White-Box / Interpretable Models (Intrinsic)
Explain Black-Box / Complex Models (Post - Hoc)
Enhance Fairness of a Model
Test Sensitivity of Predictions
Model Specific vs Model Agnostic
Model Specific: Can be applied to a single model or group of models
Model Agnostic: Can be applied to any model

Taxonomy mind-map of Machine Learning Interpretability Techniques.

Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy. 2021; 23(1):18

# LIME

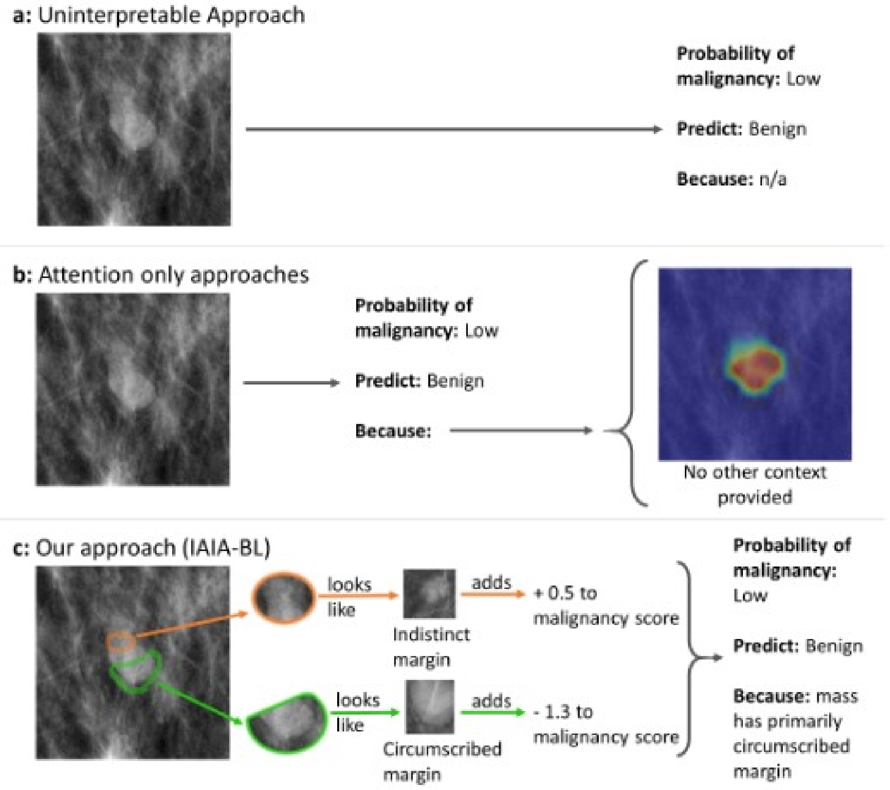
- LInear Model-agnostic Explanation
- Local method
- Algorithm :
  - Chose the instance
  - Generate new datapoints close to the instance
  - Weight the new samples w.r.t to their proximity to the instance
  - Train an interpretable (e.g. linear regression) model on the new dataset

# Shapley Values

- Based on game theory
- Feature importance among participating feature in a prediction
- For a prediction, one feature of interest:
  - Remove one feature
  - Change the value of your feature of interest
  - Make a new prediction and compute the difference
  - Repeat for all coalition of feature and take the average

# Saliency Map

- “Heat Map”
- What part of the image is important for classification ?
- Class Activation Map (CAM)

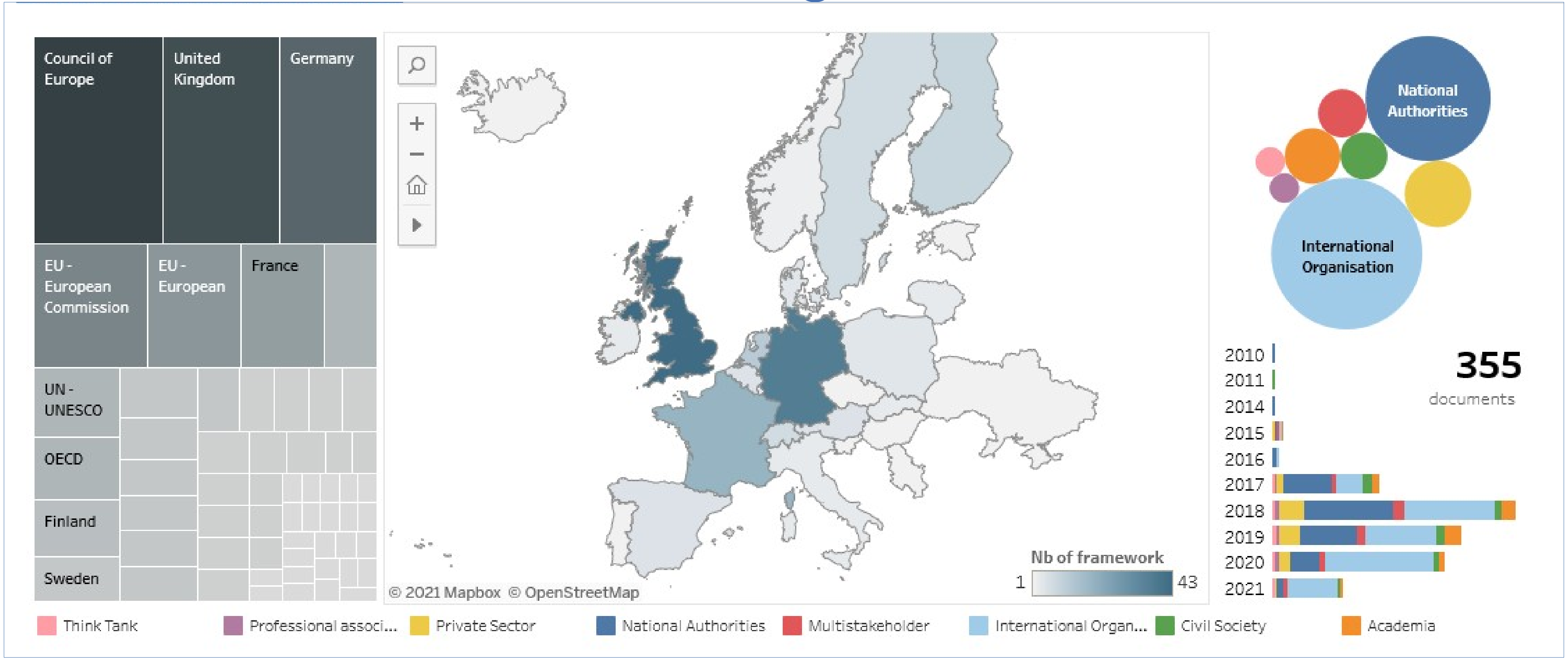

a: Uninterpretable Approach
Probability of malignancy: Low
Predict: Benign
Because: n/a
b: Attention only approaches
Probability of malignancy: Low
Predict: Benign
Because:
No other context provided
c: Our approach (IAIA-BL)
looks like
Indistinct margin
adds
+ 0.5 to malignancy score
looks like
Circumscribed margin
adds
- 1.3 to malignancy score
Probability of malignancy: Low
Predict: Benign
Because: mass has primarily circumscribed margin

Barnett, Alina Jade, et al. "A case-based interpretable deep learning model for classification of mass lesions in digital mammography." *Nature Machine Intelligence* 3.12 (2021): 1061-1070.

# Relevant Legal Texts

- GDPR : Data collection & “right to explanation”
- Medical Device Regulation: software as medical device
- Future AI Act ?
    - Does not “single out people in a discriminatory of otherwise incorrect or unjust manner”
    - Risk-based approach: MDR → High-risk
    - Provision for data collection for audot purposes

# Relevant Legal Texts



*Data Visualization of AI Initiatives from https://www.coe.int/en/web/artificial-intelligence/national-initiatives*

# Relevant Legal Texts

- GDPR : Data collection & “right to explanation”
- Medical Device Regulation: software as medical device
- Future AI Act ?
  - Does not “single out people in a discriminatory of otherwise incorrect or unjust manner”
  - Risk-based approach: MDR → High-risk
  - Provision for data collection for audit purposes

# Relevant Legal Texts: Protected Attributes in the EU

- Directive 2000/43/EC against discrimination on grounds of race and ethnic origin.
- Diretive 2000/78/EC against discrimination at work on grounds of religion or belief, disability, age or sexual orientation.
- Directive 2006/54/EC equal treatment for men and women in matters of employment and occupation.
- Directive 2004/113/EC equal treatment for men and women in the access to and supply of goods and services.
- Directive Proposal (COM(2008)462) against discrimination based on age, disability, sexual orientation and religion or belief beyond the workplace.

# Art. 9 GDPR : Processing of special categories of personal data

- 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

# Notions of Fairness

- Different Definitions, Different Metrics
  - Group Fairness, Individual Fairness
  - Demographic parity, Confusion-matrix based metrics, counterfactual fairness, Theil Index, ...
  - Sometimes incompatible
- Common ways to measure fairness: difference in accuracy and f1-score, demographic parity

# Pre-processing techniques

- Data used during training
- Collect better datasets
- Selective sampling of existing datasets
- Generative Adversarial Networks (GANs) and FairGANs
- Blindings

# In-processing techniques

- Constraint optimization
- Adversarial techniques
- Disentangled representation
- Transfer learning

# Post-processing techniques

- Reject option classification
- Equalized odds
- Calibrated equalized odds

# Relevant Conferences

- ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)
- ACM Conference on Human Factors in Computing Systems (ACM CHI)
- Computer Vision and Pattern Recognition Conference (CVPR)
- International Conference on Pattern Recognition (ICPR)
- Artificial Intelligence, Ethics and Society Conference (AIES)