Ethics Guidelines for Trustworthy AI

Course on Computer Ethics, prof. Viola Schiaffonati
Politecnico di Milano, 7 November 2019

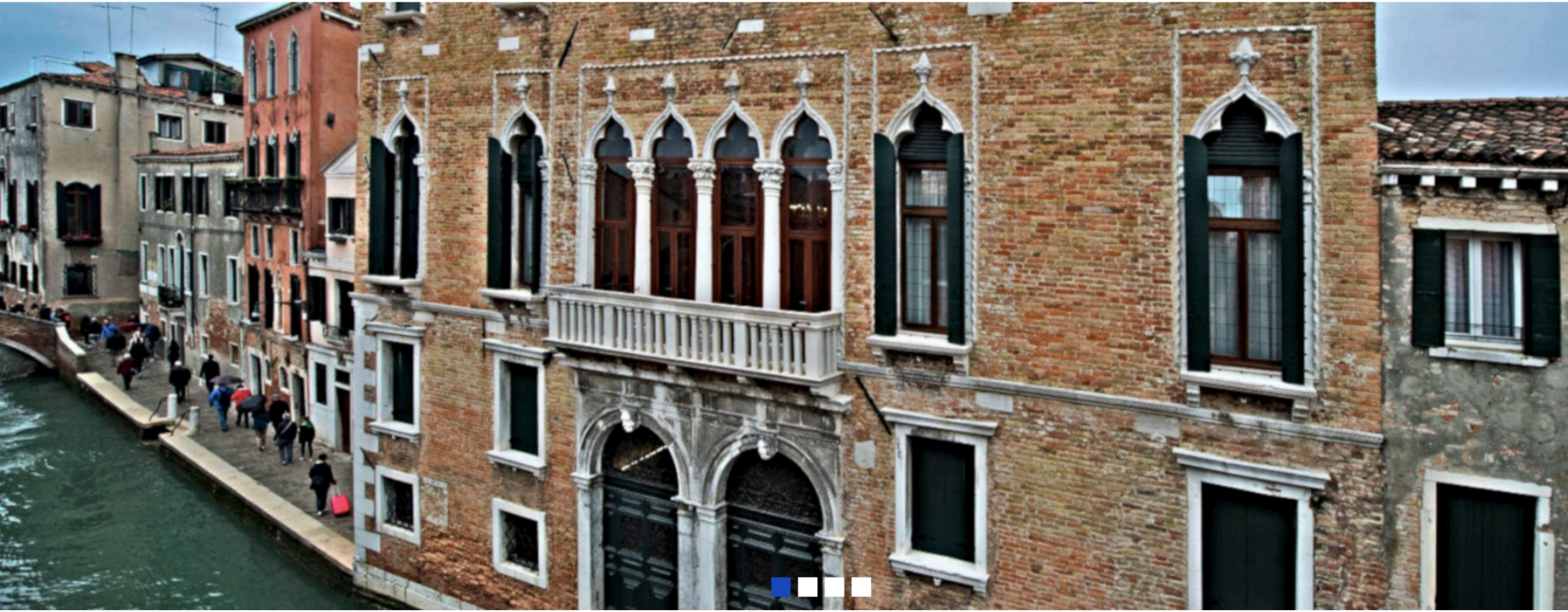
Teresa Scantamburlo

European Centre for Living Technology (ECLT)

Ca' Foscari University of Venice

Ca' Foscari University of Venice

European Centre for Living Technology

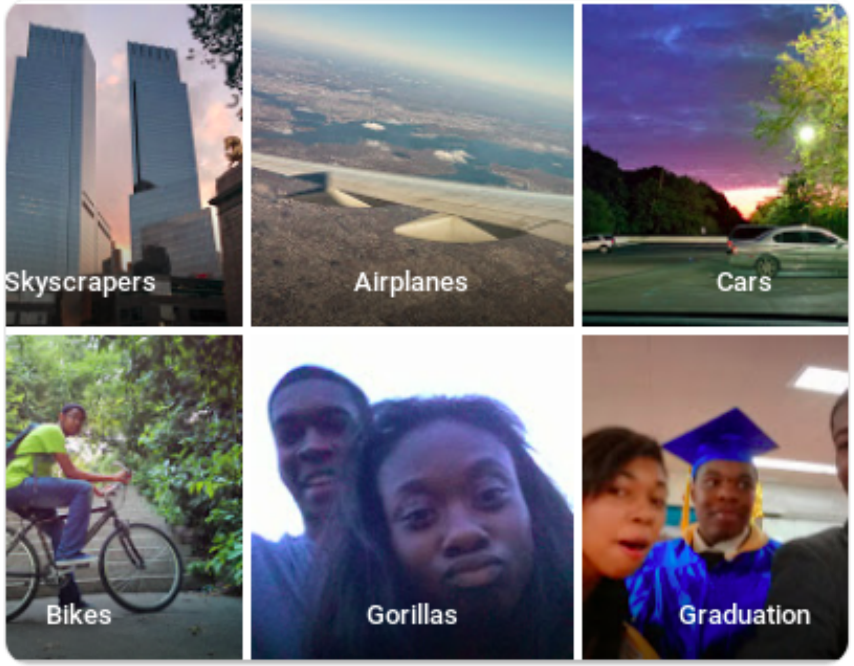


AI scandals

jackyalciné (he/him/his)
@jackyalcine

Follow

Google Photos, y'all fucked up. My friend's not a gorilla.



6:22 pm - 28 Jun 2015

3,261 Retweets **2,384** Likes

238 3.3K 2.4K

jackyalciné (he/him/his)
@jackyalcine

People-centric software consultant. 💼
black.af + koype.net; fmr @lob, @lyft, @getclef ➡️ jacky.wtf 🥬 vegan

jacky.is@black.af

Joined June 2009

Amazon built an AI tool to hire people but had to shut it do[w]
because it was discriminating against women

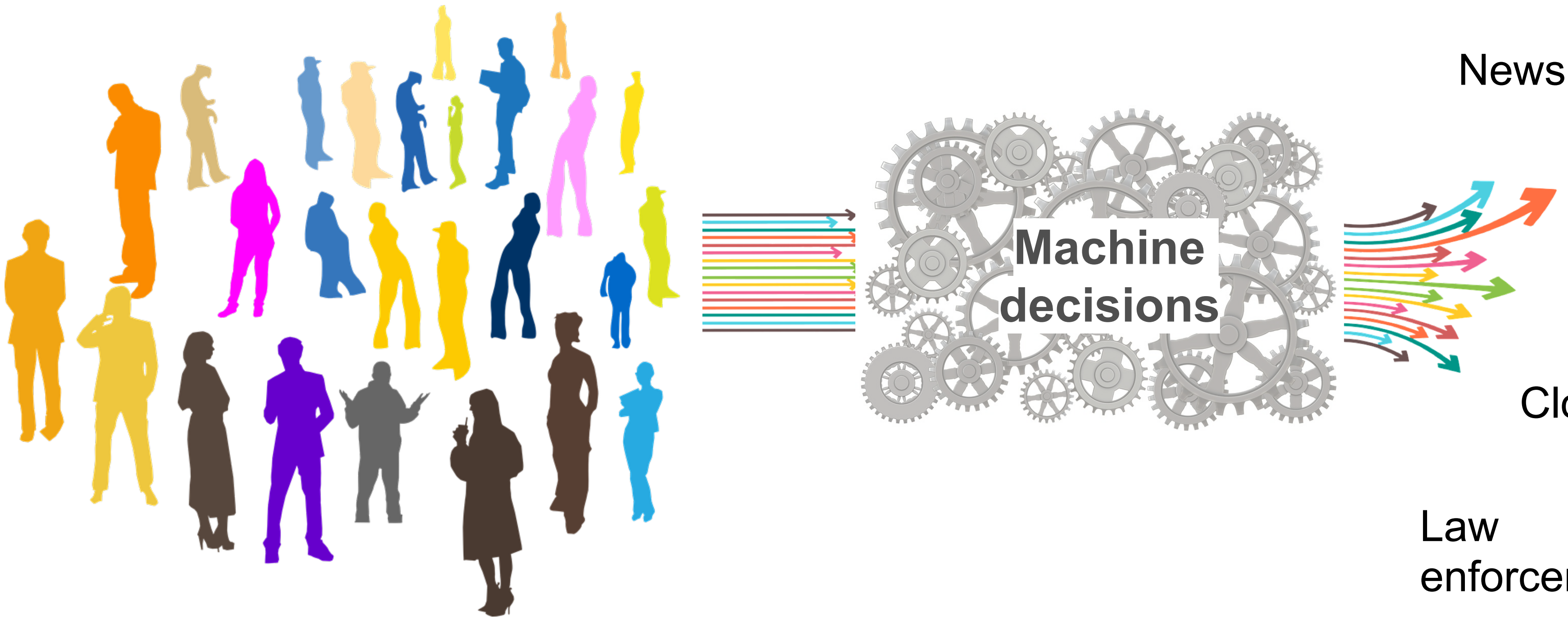
Isobel Asher Hamilton Oct. 10, 2018, 5:47 AM

The New York Times

YouTube's Product Chief on Online Radicalization and Algorithmic Rabbit Holes

Neal Mohan discusses the streaming site's recommendation engine, which has become a growing liability amid accusations that it steers users to increasingly extreme content.

Algorithms as new gatekeepers



School

News

University

Credit

Social care

Books

Health care

Clothes

Music

Job

Law
enforcement

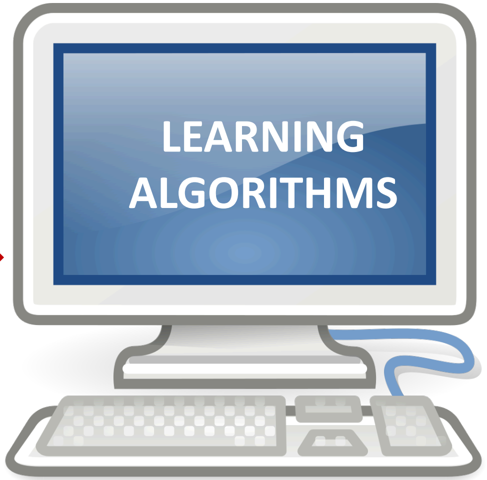
Videos

Food

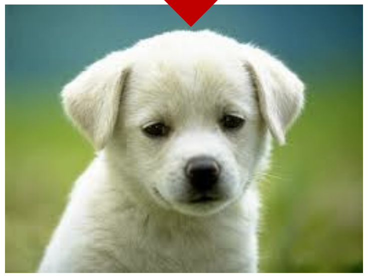
Movies

Machine Learning & diagnostic decisions

Training set



this is a dog!



? **Test set**

Can we trust AI decisions?

Outline

AI & ethics in Europe:

- Case study in law enforcement
- European commitment to Human-centred AI
- Ethics Guidelines for Trustworthy AI by the High-level Expert Group on AI
- Concluding remarks

HART: Harm Assessment Risk Tool

BBC | Sign in | News | Sport | Reel | Worklife | Travel | Future

NEWS

Home | Video | World | UK | Business | Tech | Science | Stories | Entertainment & Arts

Technology

Durham Police AI to help with custody decisions

By Chris Baraniuk
Technology reporter

10 May 2017

Share

What is HART?

- HART = Harm Risk Assessment Tool
- It is a Risk Assessment Tool (RAT) that is used to predict the likelihood of reoffending after a follow-up period (i.e. 2 years after arrest)
- RATs are usually based on statistics or machine learning
- They can be introduced at several steps of the justice process, e.g. pre-trial hearing, early release from prison (parole), sentencing, etc.
- There are several RATs in use both in US and Europe, e.g.:
  - USA: COMPAS, Public Safety Assessment Tool, Ohio Risk Assessment System… (for a list see: https://epic.org/algorithmic-transparency/crim-justice/)
  - Europe: HART (England), OGRS (England and Wales), StatRec (Netherland), Static99 (just for sexual offenders, Netherland)

Our sources

- The analysis of HART is based on the following sources:
  - Urwin S (2016) *Algorithmic Forecasting of Offender Dangerousness for Police Custody Officers: An Assessment of Accuracy for the Durham Constabulary*. Master Degree Thesis, Cambridge University, UK
  - Oswald M, Grace J, Urwin S and Barnes GC (2018) Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality, *Information & Communications Technology Law*, 27(2): 223-250
  - Barnes G, and Hyatt J (2012) , Classifying Adult Probationers by Forecasting, Future Offending, Tech report
- Extensive media coverage
  - BBC: https://www.bbc.com/news/technology-39857645

A brief sketch

- Launched in May 2017
- Developed by Durham Police in collaboration with Cambridge University
- **Objective**: to support custody decision = "decision taken by the custody officer following arrest at the end of the first custody period" (Urwin, 2016)
- Model's **output**: "high-risk" – "moderate-risk" – "low-risk"
- Context: "checkpoint programme" that aims at providing "moderate-risk offender" with an alternative to prosecution (https://www.durham.police.uk/Information-and-advice/Pages/Checkpoint.aspx)

HART's training set

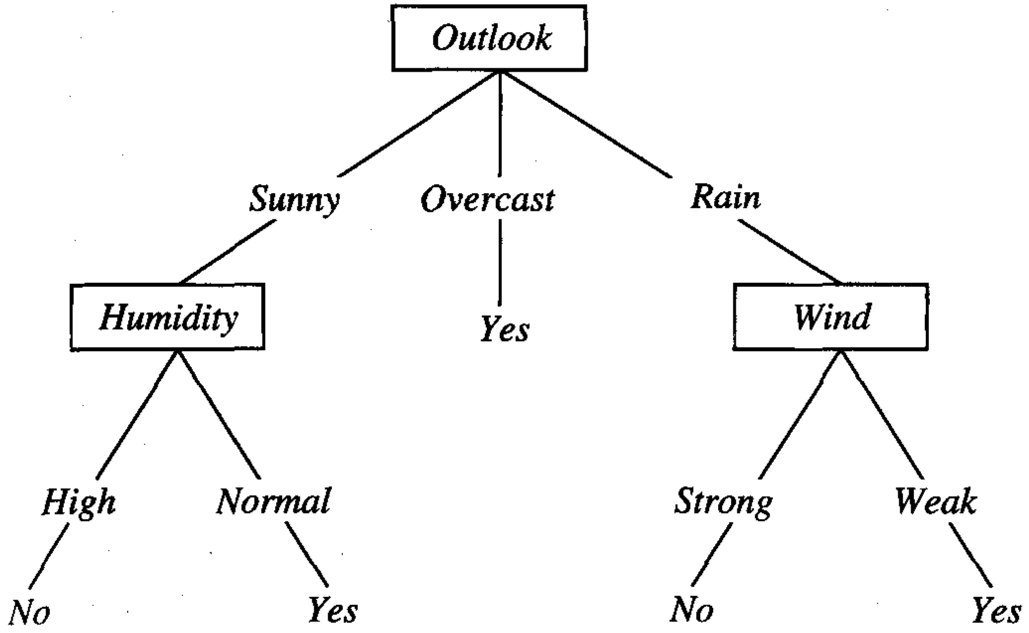
- 104,000 custody events within a period between Jan 2008 and Dec 2012

- 34 features such as:
  - Age at custody event
  - Gender
  - Count of any past offences
  - Instant violence offence (Y/N)
  - Custody Outward Postcode (3-4 first characters)
  - (Experian) Mosaic Code (socio-geo demographic)
  - Age at first offence
  - …

- Categorical labels:
  - High-risk = a new serious offence within the next 2 years
  - Moderate-risk = a non-serious offence within the next 2 years
  - Low-risk = no offence within the next 2 years

Decision tree

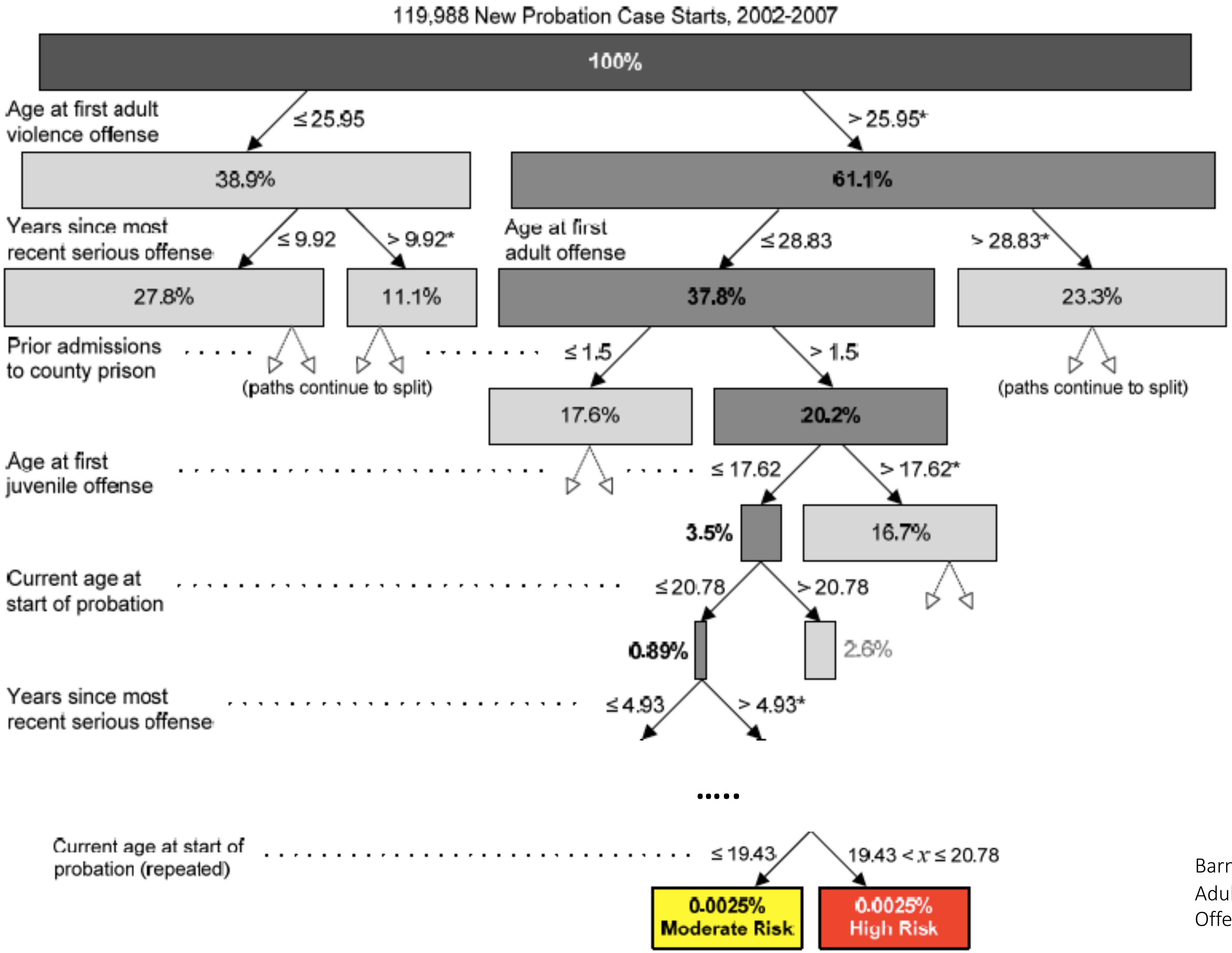
- A decision tree is a popular classification technique that tests an attribute at each node and assign instances to the descending branches based on the value taken by instances for that attribute



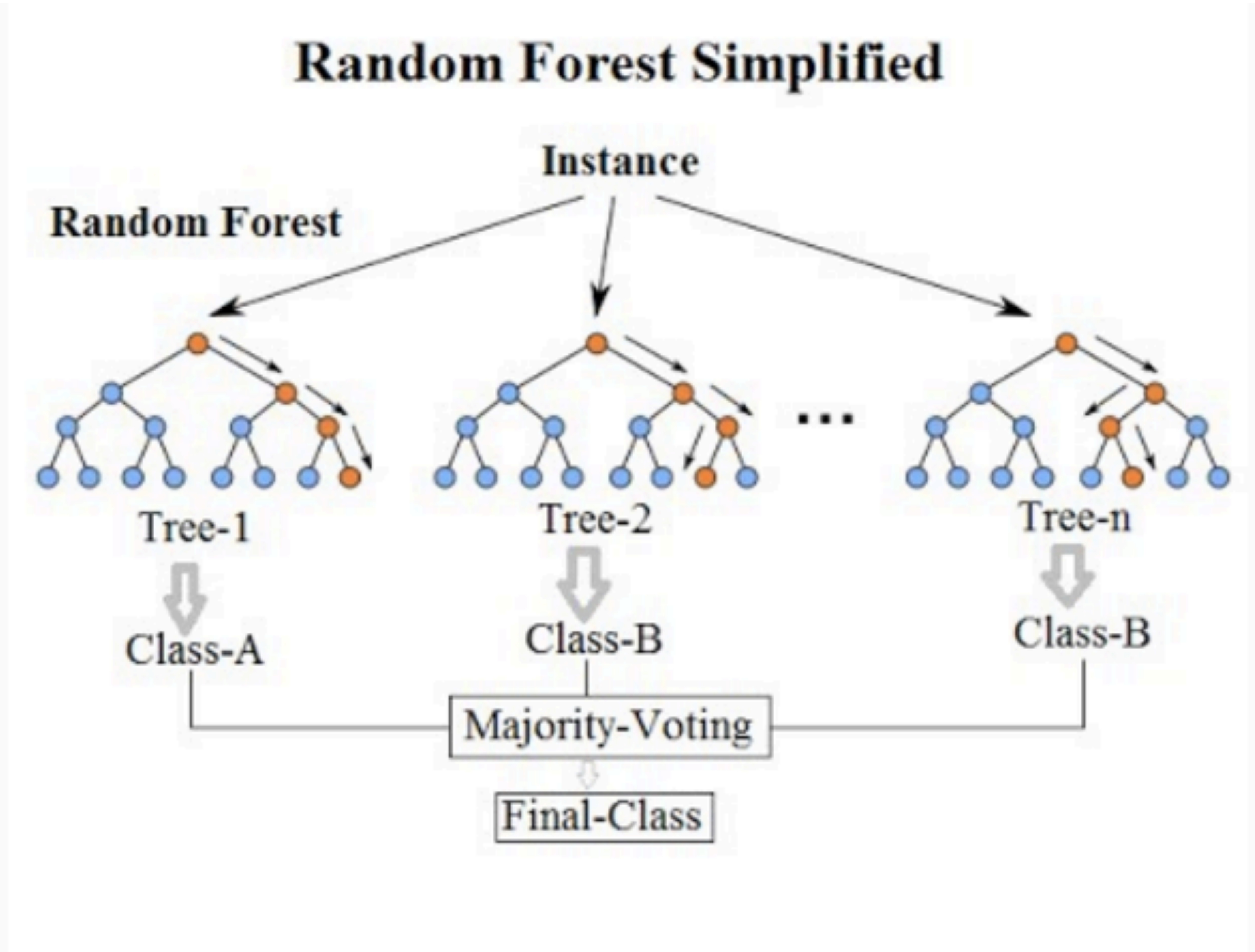
Simple example of a decision tree from Mitchell T, *Machine Learning*, 1997

HART’s model

- HART is based on **Random Forest**, a ML method that results from the combination of a multitude of decision trees
- Each decision tree is trained on a **random subsamples** of the training set and using a **random subset** of **features**
- HART uses **509 decision trees**, each producing a prediction. The output corresponds to the output that receives the most votes



Barnes G, and Hyatt J (2012) , Classifying Adult Probationers by Forecasting, Future Offending, Tech report



https://commons.wikimedia.org/wiki/File:Random_forest_diagram_complete.png

Confusion matrix

2013 Validation	Actual High	Actual Moderate	Actual Low	Total
Forecast High	6.26%	10.01%	2.23%	18.49%
Forecast Moderate	4.88%	32.53%	13.55%	50.95%
Forecast Low	0.73%	5.81%	24.02%	30.55%
Total	11.86%	48.35%	39.79%	100% = 14,882 custody events

Confusion Matrix for the test set (see table 8 in Urwin, 2016: 54)

Technical assessment

- Out-of-bag error = when a random sample is drawn to grow a decision tree a small amount is held out and used as a test set to estimate the generalization error during training

- Weighting different types of errors:
  - **Dangerous errors**: misclassifying a serious offender as a low-risk
  - **Cautious errors**: misclassifying a non-serious offender as a high-risk

- Policy decision: HART weights more dangerous error (i.e. it applies a lower cost-ratio)

Performance measures

Comparison with the accuracy of a random baseline:

[P(Y = “high”) * P(Ŷ= “high”)] + [P(Y = “moderate”) * P(Ŷ= “moderate”)] + [P(Y = “low”) * P(Ŷ= “low”)] =

[0.1186 * 0.1186] + [0.4835 * 0.4835] + [0.3979 * 0.3979] = 0.406 = **41%**

	OOB construction data	2013 validation data	
Overall accuracy: what is the estimated probability of a correct classification?	68.50%	62.80%	
Sensitivity / recall: what is the true positive rate for each class label?	72.60%	52.75%	HIGH
	70.20%	67.28%	MODERATE
	65.30%	60.35%	LOW
Precision: what is the rate of relevant instance for each class label?	48.50%	33.83%	HIGH
	70.20%	63.84%	MODERATE
	75.60%	78.60%	LOW
Very dangerous errors: of those predicted low risk, the percent that was actually high risk (subset of the false omission rate)	2.40%	2.38%	
Very cautious errors: of those predicted high risk, the percent that was actually low risk (subset of the false discovery rate)	10.80%	12.06%	

some performance measures of HART extracted from tables 6 and 9 in Urwin (2016: 52,56)

From accuracy to trust

- Being accurate is not enough
  - What performance measures are used?
  - What sample is used for validation?
- Can results be reproduced?
- How are decisions made?
- What model has been used? What features?
- Can we explain the logics behind the algorithm to the interested subject? (GDPR)
- Is the algorithm fair? Or does it discriminate?
- How does the user (police officer / judges / doctors..) approach machine learning predictions?
- ...

What is Europe doing?

European approach to AI

“Artificial Intelligence for Europe” COM(2018) 237, 25 April 2018

The European initiative aims to:

- “Boost the EU's technological and industrial capacity and AI uptake across the economy”
- “Prepare for socio-economic changes brought about by AI”
- “Ensure an appropriate ethical and legal framework, based on the Union's values and in line with the Charter of Fundamental Rights of the EU”

“Coordinated Plan on Artificial Intelligence” COM(2018) 795, 7 December 2018

“Overall, the ambition is for Europe to become the world-leading region for developing and deploying cutting-edge, ethical and secure AI, promoting a human-centric approach in the global context.”

Human-centric AI

In short human-centric AI implies:

- People can trust AI systems (trustworthy AI)
- Individuals and the society can benefit from the use of AI
- AI systems are based on ethical and societal values, in particular, the European Charter of Fundamental Rights

In more concrete terms:

- ethical and secure by design
- clear ethics guidelines and standards
- legal framework

Ethics guidelines

High-level Expert Group on Artificial Intelligence (AI HLEG)

AI HLEG's main deliverables:

- AI Ethics guidelines delivered
- Policy and investment Recommendations

AI HLEG's ethics guidelines:

- first draft December 2018
- public consultation
- official delivery in April 2019
- piloting process with the support of AI4EU (June-December 2019)

Website: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Trustworthy AI

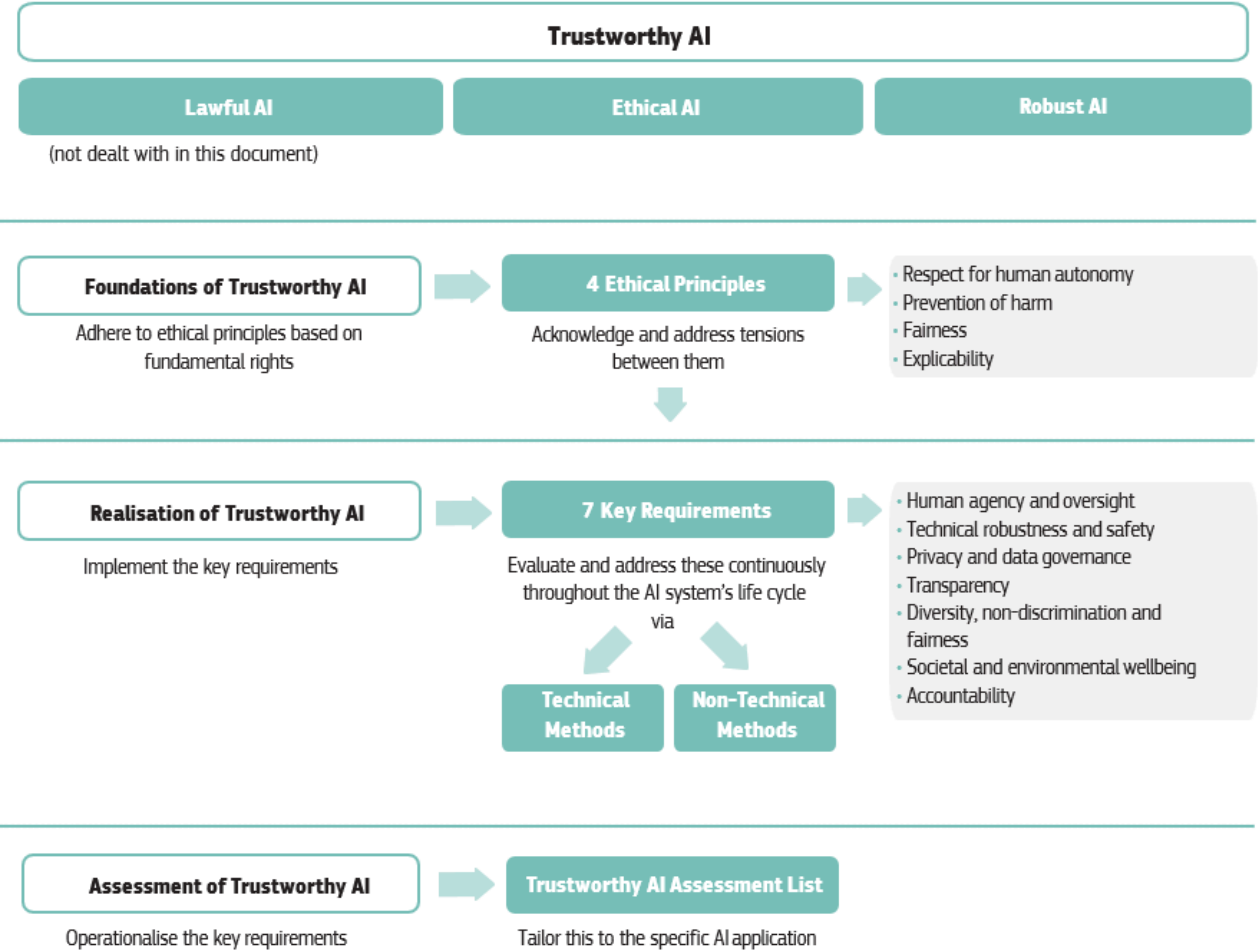
"AI systems need to be human-centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom."

Trustworthy AI (instead of ethical AI)

- being demonstrably worthy of trust (concrete pathways)
- it refers to the socio-technical system in which AI technology is embedded (holistic approach)
- Trustworthy AI to promote "responsible competitiveness"
- Addressed to AI stakeholders, e.g. companies, civil society organisations, individuals, ...

Some remarks:

- Trustworthy AI is a contribution to elaborate "a normative vision of an AI-immersed future"
- need of an ethical culture through public debate, education and practical learning



Framework

AI HLEG, *Ethics Guidelines for Trustworthy AI* (2019, p 8)

Addressing requirements

They can help the implementation of trustworthy AI

1. human agency and oversights
2. technical robustness and safety
3. privacy and data governance
4. transparency
5. diversity, non-discrimination and fairness
6. societal and environmental well-being
7. accountability

value-by-design methodology

fairness metrics

multiple performance measures

explainable AI methods

testing performed by diverse groups

adversarial testing

codes of conduct

“bug bounties”

regulation (e.g. GDPR)

accountability via governance frameworks

education

stakeholder participation

diversity and inclusive design teams

Trustworthy assessment list

Brief sketch:

- list of questions structured around the 7 requirements
- goal = to operationalise the key requirements
- primarily addressed to developers and deployers of AI systems
- compliance with this list is not evidence of legal compliance
- piloting process (qualitative and quantitative assessment)

Assessment List: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60440

Towards trustworthy AI

Some insights:

- holistic approach, being open to changes (business models)
- diversity and inclusion (design, validation, deployment)
- disseminate results and communication to the public (realistic expectations, open questions)
- long term solutions, gradual and dynamic process (ethical culture)

Some weaknesses:

- being demonstrably trustworthy is hard
- some methods for implementing requirements are too abstract
- assessment list include too many questions
- risks of applying requirements/assessment list in a mechanical way

Main references

- Scantamburlo T., Charlesworth A. and Cristianini N. (2019). "Machine decisions and human consequences". In K. Yeung & M. Lodge (eds) *Algorithmic Regulation*, Oxford: Oxford University Press (the draft accepted for publication is available on arXiv)

- High-Level Expert Group on Artificial Intelligence (2019), Ethics Guidelines for Trustworthy AI, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Thanks for your attention

Feedbacks, comments or requests are welcome

teresa.scantamburlo@unive.it