RESPONSIBLE MACHINE LEARNING

Monowar Bhuyan

Department of Computing Science

Umeå University, Sweden

https://people.cs.umu.se/monowar/

RML Winter School Mar 11-13, 2025, Umeå

UMEÅ UNIVERSITY

11 March 2025

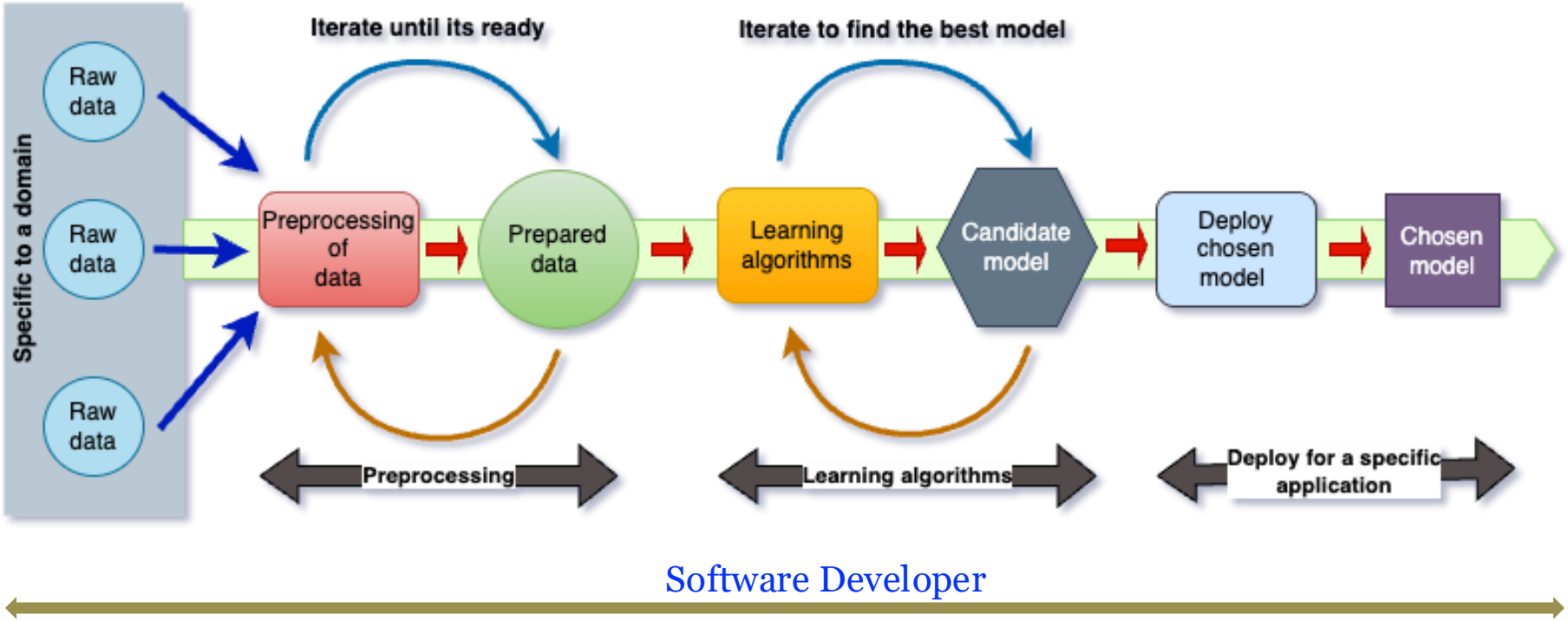
OUTLINE

- Introduction
- Responsible ML
  - Principles of RML
  - Development process
  - Open challenges
- Conclusion

INTRODUCTION

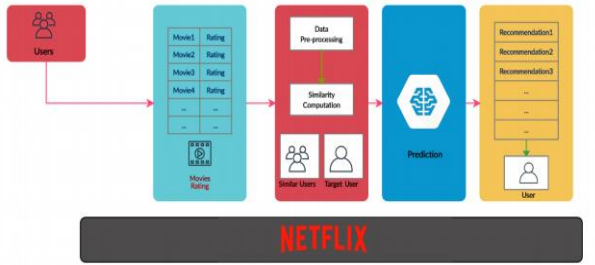
- AI/ML design and deployment lack consideration of multi-stakeholder's input, values and norms to make a sustainable and safe society
- Societal impacts get either positive or negative depending on the context
- Hard to consider everything from each stakeholder during the design, implementation, and deployment of software for day-to-day usage
- Machine learning developments mostly focused on data, model architecture and performance

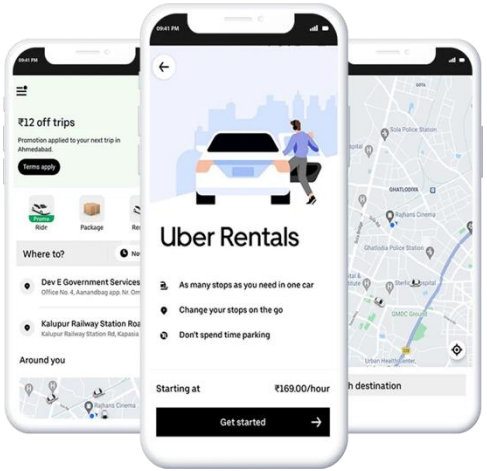
ML DEVELOPMENT PROCESS

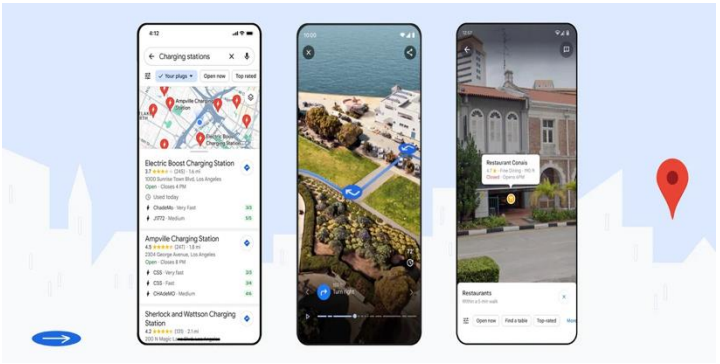


AI USES IN DAY-TO-DAY LIFE

- Software - voice assistants, image recognition for face unlock in mobile phones, and ML-based financial fraud detection
- Embodied - drones, self-driven vehicles, assembly-line robots, and the Internet of Things (IoT) devices
- Examples:







Source: https://insights.daffodilsw.com/blog/20-uses-of-artificial-intelligence-in-day-to-day-life

ADVANCES IN ML: CONCERNS AND THREATS

- Fairness: Machine learning models often discriminate against individuals or groups based on protected characteristics such as race, gender, age, religion, or other attributes.

- "ChatGPT and -by extension- LLMs (if not properly monitored) could be propagators and amplifiers of negative or discriminatory stereotypes related to social or ethnic groups or religious, political, and even sexual orientations." *[Hartvigsen et al., 2022]*

Amazon discontinued a recruiting algorithm after discovering that it led to gender bias in its hiring. (Credit: Brian Snyder/Reuters)

DALL·E 3

[1]

ChatGPT

Stable Diffusion

[1] image: https://www.studyiq.com/articles/generative-ai/

BRIDGING THE GAP: RAI AND RML

Responsible AI

- Principles to practice
- Assesses existing systems according to the principles
- RAI also limits to the modelling level rather than investigating

Responsible ML

- Integration of ML development process with RAI
- Assesses and measures societal impacts
- Integrating multi-stakeholders input

Frameworks	Societal context	Ethical, legal, and societal requirements	Data	Models	Evaluating impact
ML pipeline	✗	✗	✓	✓	✗
Responsible AI	✓	✓	✗	✗	✓
Responsible ML	✓	✓	✓	✓	✓

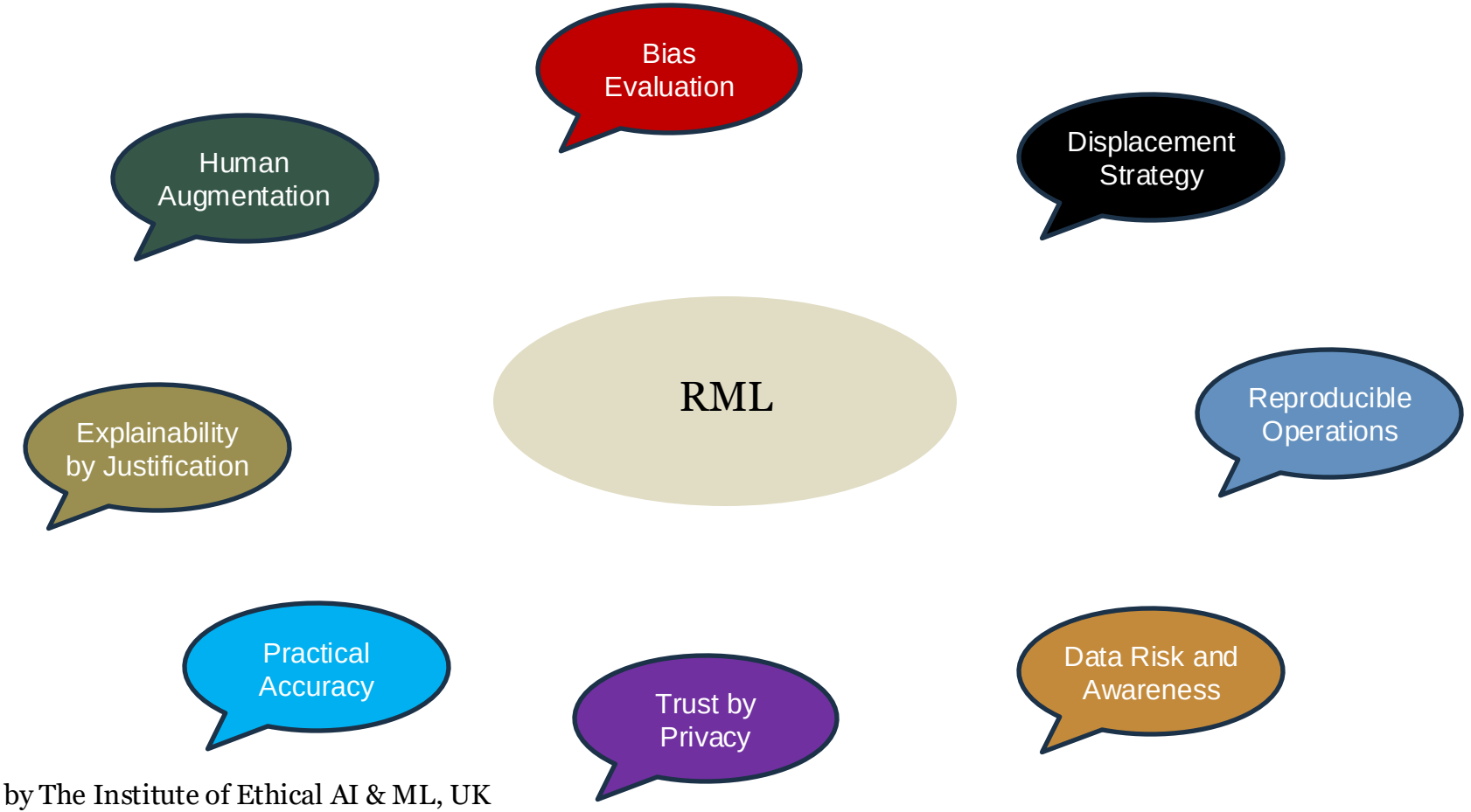
RESPONSIBLE ML

- Responsible ML practices are required to better understand, protect, and control data, models, societal impacts, and processes to build trusted solutions.

- The concept of responsible ML needs time to evolve and grow with input from:
  - Diverse practitioners
  - Researchers
  - Decision makers/policymakers
  - Users

- *Def*: Responsible ML implements ethical, legal, and societal requirements into the ML pipeline and justifies and evaluates the ethical trade-offs that arise in this pipeline in order to address the societal impact of ML systems

8 PRINCIPLES OF RESPONSIBLE ML

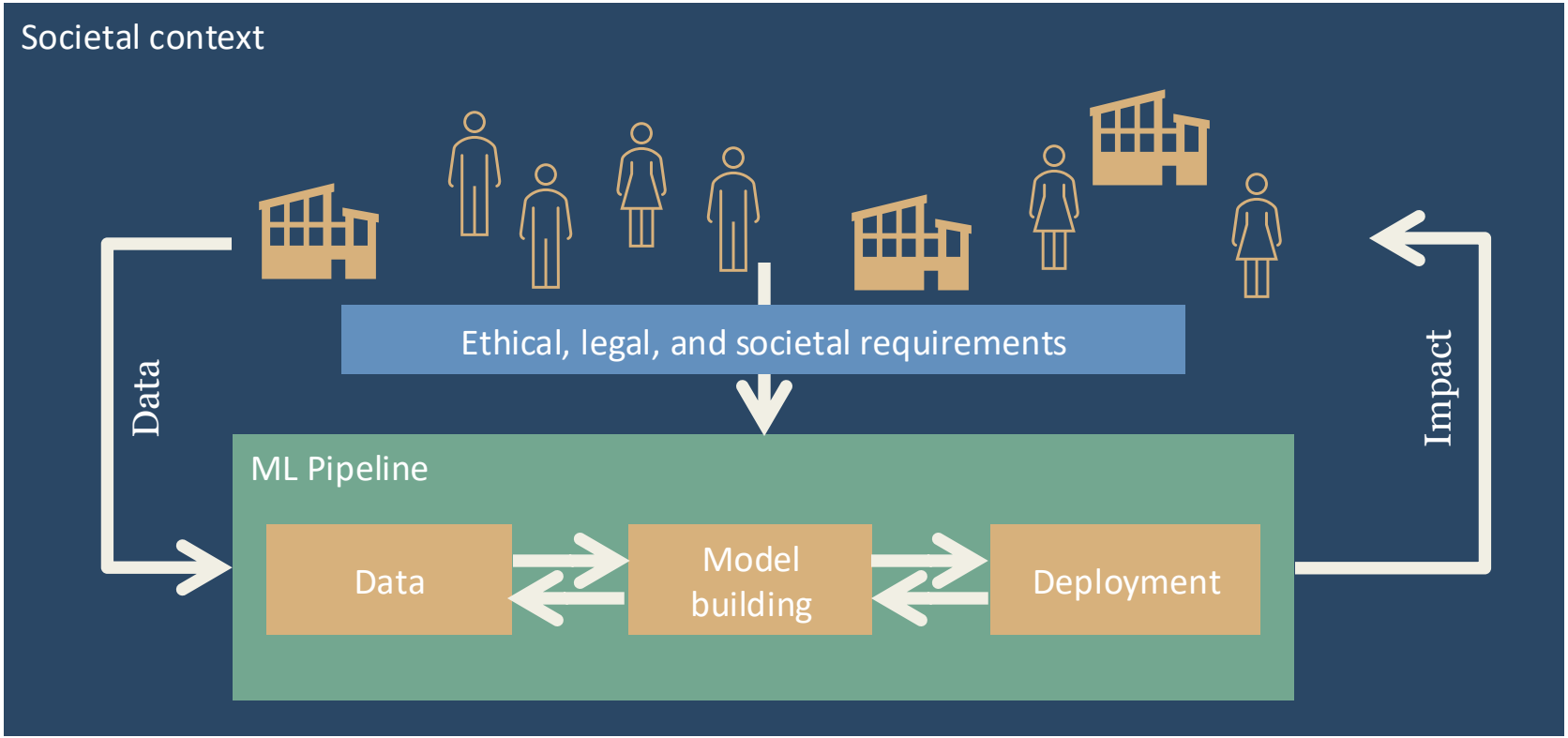


Provided by The Institute of Ethical AI & ML, UK

8 PRINCIPLES OF RESPONSIBLE ML

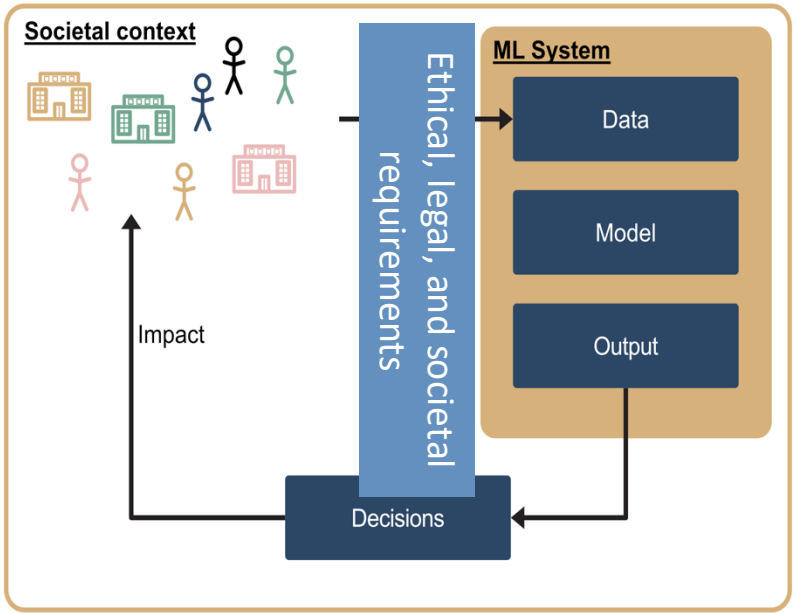
- Human augmentation
  - Human in/on the loop process
  - E.g., credit card fraud detection
- Reproducible operations
  - Develop infrastructure that enables reproducibility
  - E.g., abstracting each computational step (model reproducibility)
- Displacement strategy
  - Processes to reduce impact, such as adaptability
  - E.g., move from one organization to another
- Practical accuracy
  - Accuracy and cost metric functions are aligned to the domain-specific applications
  - E.g., Domain-specific metrics

RESPONSIBLE ML FRAMEWORK

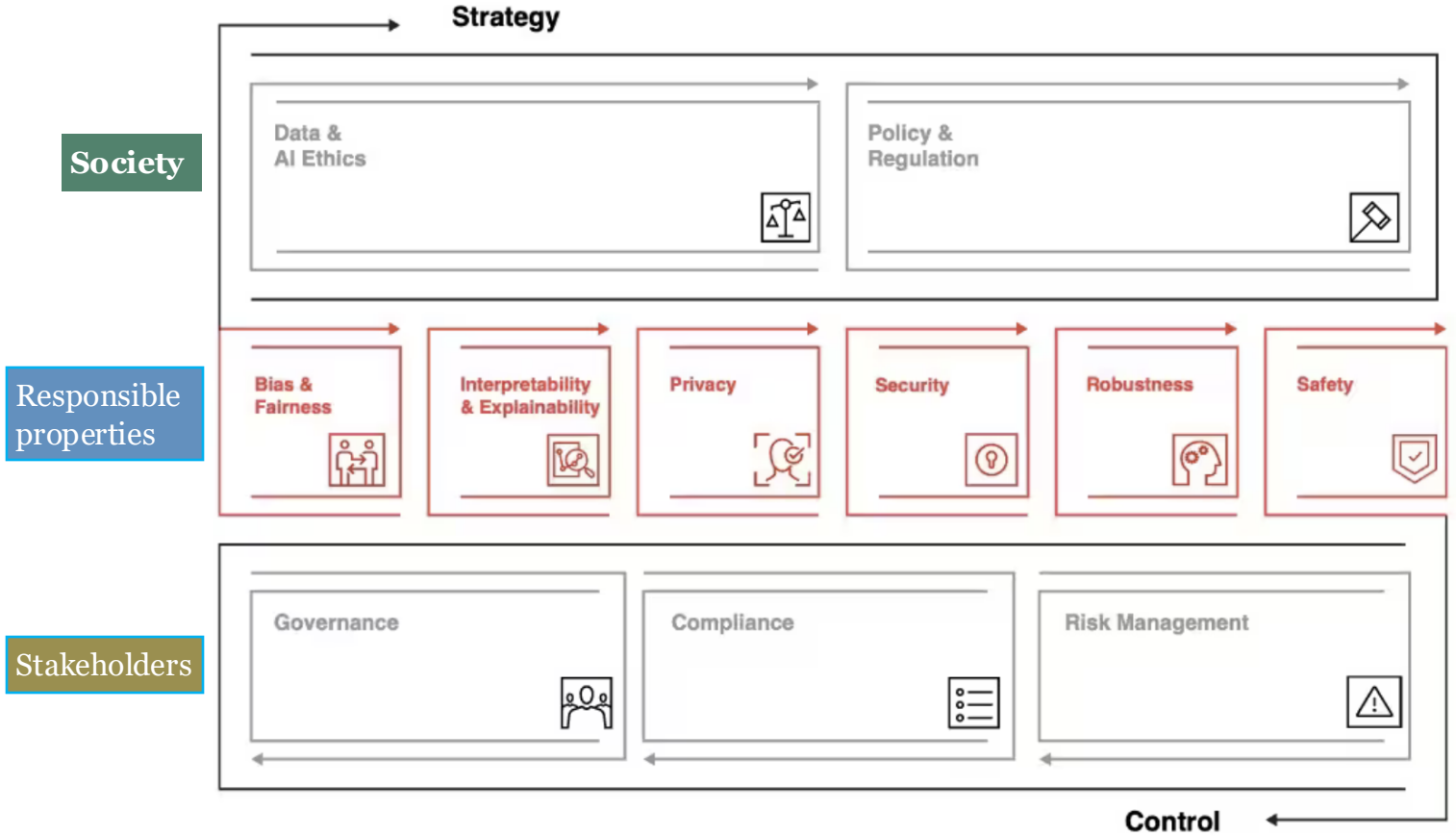


RESPONSIBLE ML FRAMEWORK

- Incorporates ML development process in Responsible AI life cycle
- Analysis of available data
  - Identify potential issues
- Design requirements for
  - Data
  - Model
  - System
- Verification of designed systems
  - According to responsible properties (e.g., fairness, privacy)



DIMENSIONS OF RML



Source: PwC

MAKING TRADE-OFFS

- Stakeholders will have contradicting ideas about properties
  - And will find different properties more important

- Need to make trade-offs between those ideas in the design
  - **For example**: there are multiple ways to make a system fair

- Trade-offs in implementation
  - After coming up with design requirements for each property
  - How will you balance among the different properties?
    - More fairness means less accuracy

- There is no perfect solution, but you need to make your choices transparently.
  - Who made the choice?
  - How did you make the choice? What options were considered?
  - Why did they pick the option they picked?

FAIRNESS

UMEÅ UNIVERSITY

FAIRNESS IN ALGORITHMIC DECISION-MAKING

- **Data**
  - Sensitive attributes (e.g., gender)
  - Non-sensitive attributes (e.g., high school grades)
  - Label/ground-truth (e.g., university grades)

- **Algorithmic** decision-making
  - Policy/predictor predicts label/ground-truth (e.g., graduation) to make decisions (e.g., university admission)

STATISTICAL FAIRNESS – LIMITATIONS

- Individual Fairness
  - *Idea*: treating similar individuals similar
  - Difficulty: defining a similarity function

- Group Fairness
  - *Idea*: treating demographic groups on average similar
  - Difficulty: capturing discrimination without, for instance, a "causal story", that defines groups

BERKELEY ADMISSIONS SCENARIO

Men		Women	
Applied	Admitted (%)	Applied	Admitted (%)
8442	44	4321	35

Evidence of discrimination?

BERKELEY ADMISSIONS SCENARIO

	Men		Women	
Department	Applied	Admitted (%)	Applied	Admitted (%)
A	825	62	108	82
B	520	60	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Indirect discrimination because the Department made choices for admission

OPEN PROBLEMS

- Getting the right data, model development, and deployment
  - Integration of responsible properties
  - Observe direct and indirect impacts on society
- Fairness sensitivity analysis for learning with multiple representations (e.g., text, image, audio, video)
- Assess transparency from design to deployment
- Accountability of data and ML models, for example, under adversarial manipulations
- Sandboxing and ethical implementation platform
- Explainability of data, model and decision concerning a context

CONCLUSION

Are we all responsible?

- Nothing fits in one solution
- Human action and intention is a crucial underpinning of responsible innovation

Take away

- The design and implementation of algorithmic models as an eminently human activity— an activity guided by our purposes and values, an activity for which each of us who is involved in the development and deployment of AI systems is morally and socially responsible.

- Alan Turing Instiute

REFERENCES

1. Virginia Dignum. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (1st. ed.), Springer.

2. Barocas, Solon, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

3. Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. 2023. *"How Biased are Your Features?": Computing Fairness Influence Functions with Global Sensitivity Analysis*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 138–148. https://doi.org/10.1145/3593013.3593983

4. Moraffah, Raha, et al. "Socially Responsible Machine Learning: A Causal Perspective." *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023.

5. Marybeth Defrance and Tijl De Bie. 2023. *Maximal fairness*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 851–880. https://doi.org/10.1145/3593013.3594048

6. Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside. "An Overview of Catastrophic AI Risks." *arXiv preprint arXiv:2306.12001* (2023).

7. Chen, Pin-Yu, and Payel Das. "AI Maintenance: A Robustness Perspective." *Computer* 56.2 (2023): 48-56.

ACKNOWLEDGMENT

MSCA Doctoral Networks - LEMUR

Funded by
the European Union

Virginia Dignum and Pim Kerkhoven

monowar@cs.umu.se