UiO : Institutt for informatikk

UiA Universitetet i Agder

AI4Users

Explainable to whom?

Graduate AI Ethics Course – 2021

Alexander Kempton (University of Oslo) and Polyxeni Vassilakopoulou (University of Agder)

Polyxeni (Xenia) Vassilakopoulou

Professor, Department of Information Systems, University of Agder.

Main research interests:

Data management and the evolution of data-intensive infrastructures

Dynamics and governance of human – AI arrangements

Design-oriented studies with a sensitivity to user perspectives

Alexander Kempton

Associate Professor, Department of Informatics, University of Oslo
Information Systems Group and HISP Center

Main research interests:

Innovative and responsible generation and use of data

Digital government and digitalization of public sector

AI4Users

Research project: a human-centred perspective for the introduction of AI in public services. The overall aim is to enhance public confidence and societal value.

Two pillars:

Enable human control: AI intelligibility

Ensure ethically aligned design: AI accountability

Research project for 4 years (till December 2024) following an Action Design Research (ADR) approach. Funded by the Norwegian Research Council.

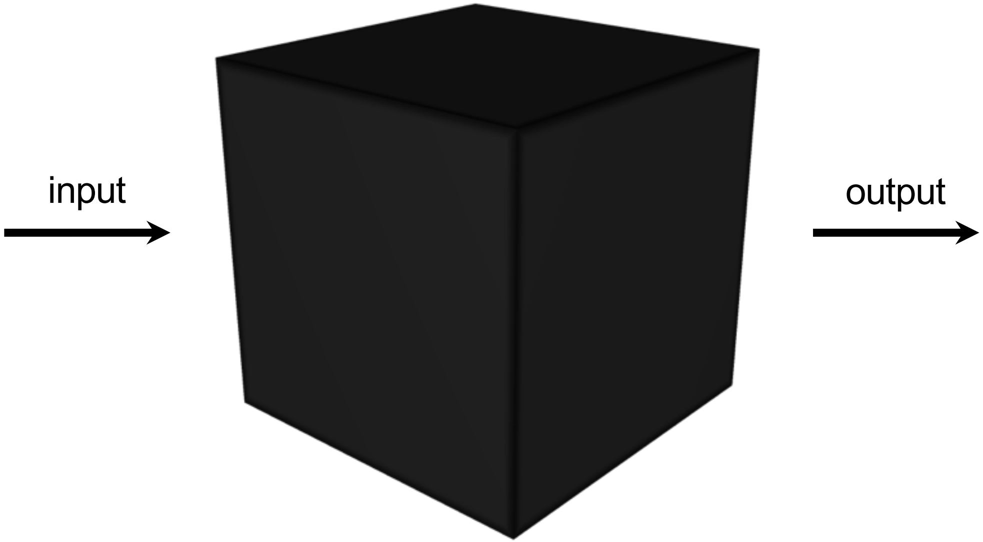
Partners:

- University of Agder (leader)
- University of Oslo
- Norwegian University of Science and Technology (NTNU)

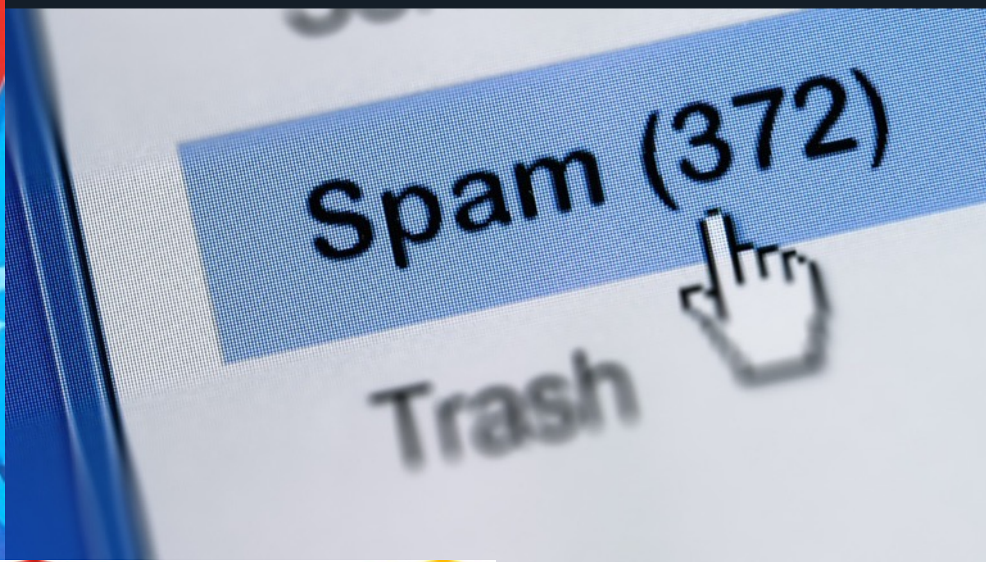
AI explanation needs for different audiences

AI explanations within work practices

“black boxes”

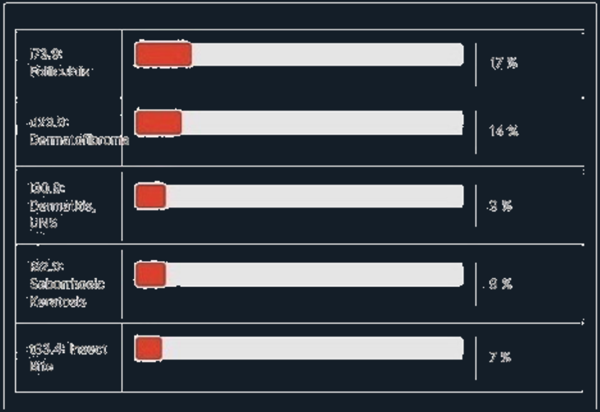


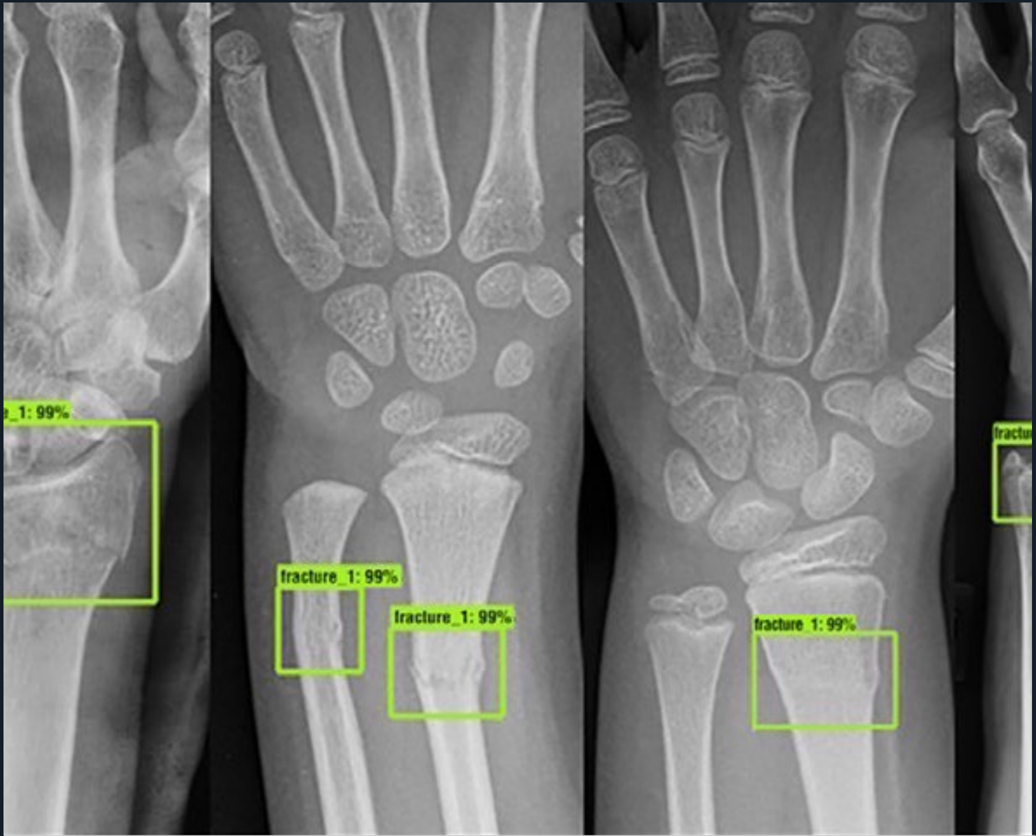
Recommendations



Spam Detection

Pattern Detection





Radiology Diagnostics

AI in public services

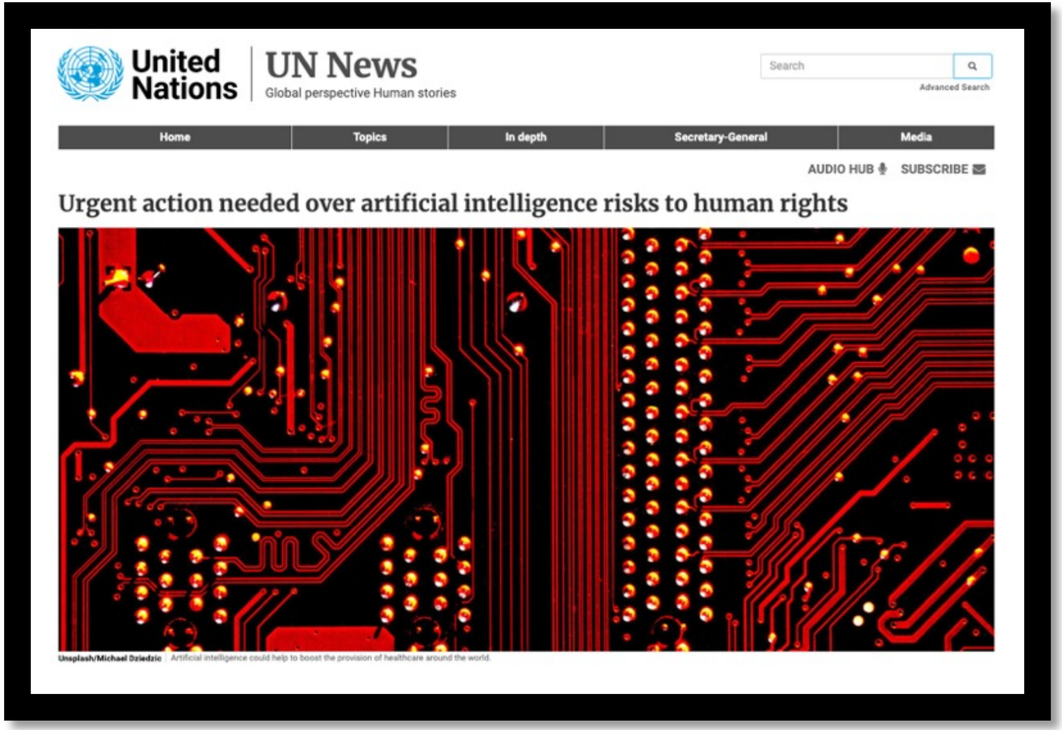
Needs for explanation in different AI application domains

https://padlet.com/polyxenivassilakopoulou/explnationsneed

“

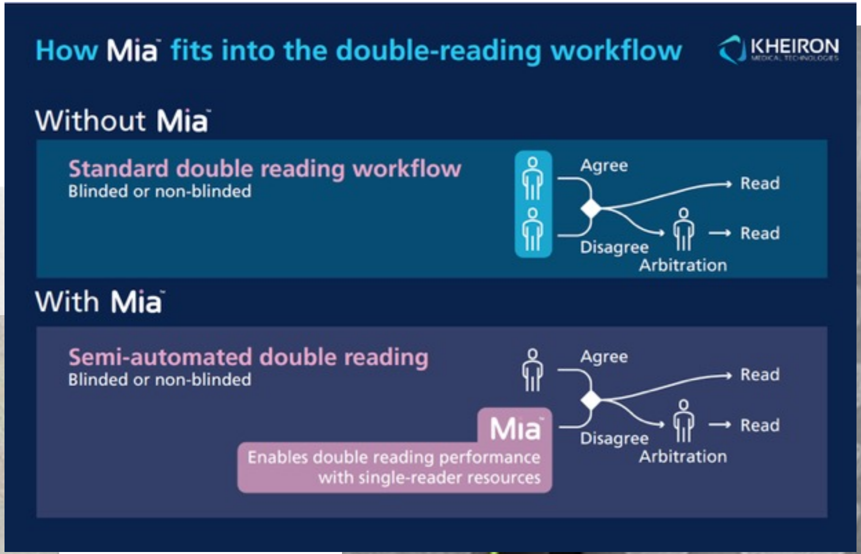
“

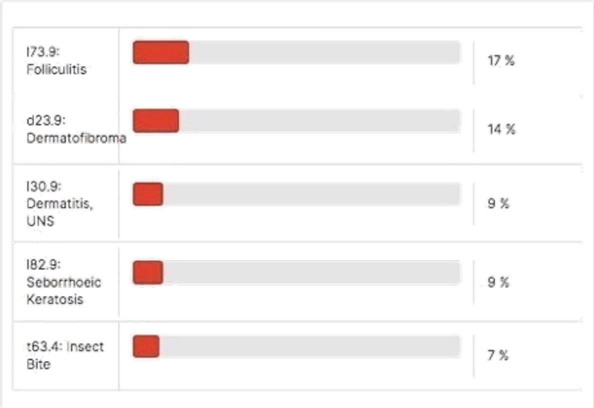
SyRI is a risk estimation model introduced in the Netherlands to assess individuals' likelihood for benefit fraud. In February 2020 the District Court of the Hague ordered its halt **due to its opaqueness** and lack of sufficient safeguarding mechanisms to protect privacy.

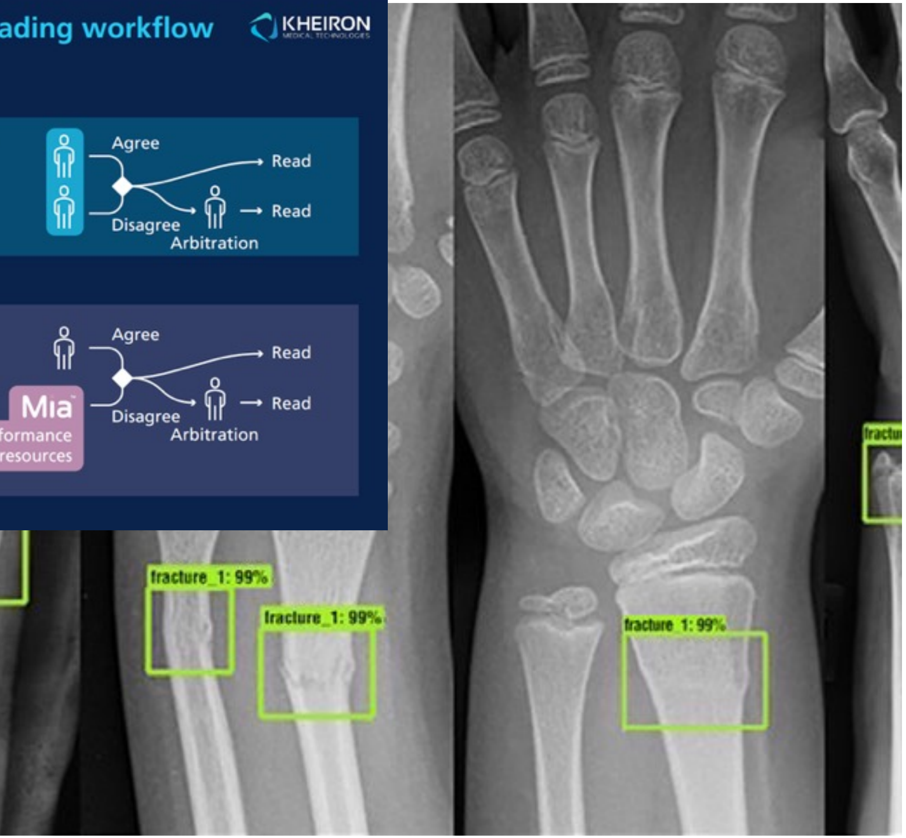


Requested filling the immense accountability gap and **requesting greater transparency**

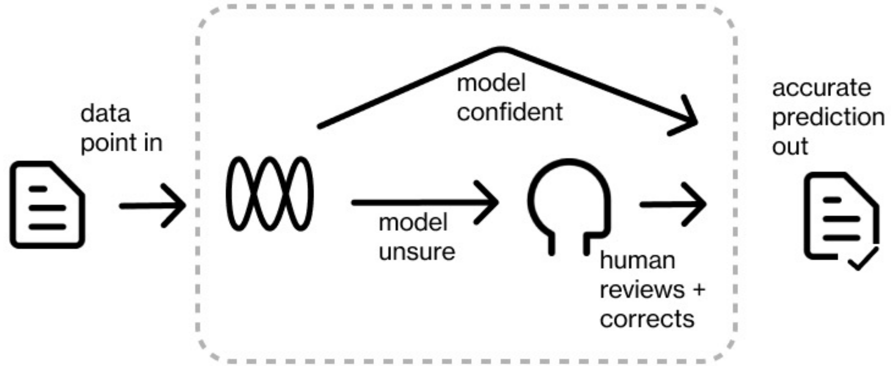
Pattern Detection







Radiology Diagnostics


data
point in
model
confident
model
unsure
human
reviews +
corrects
accurate
prediction
out

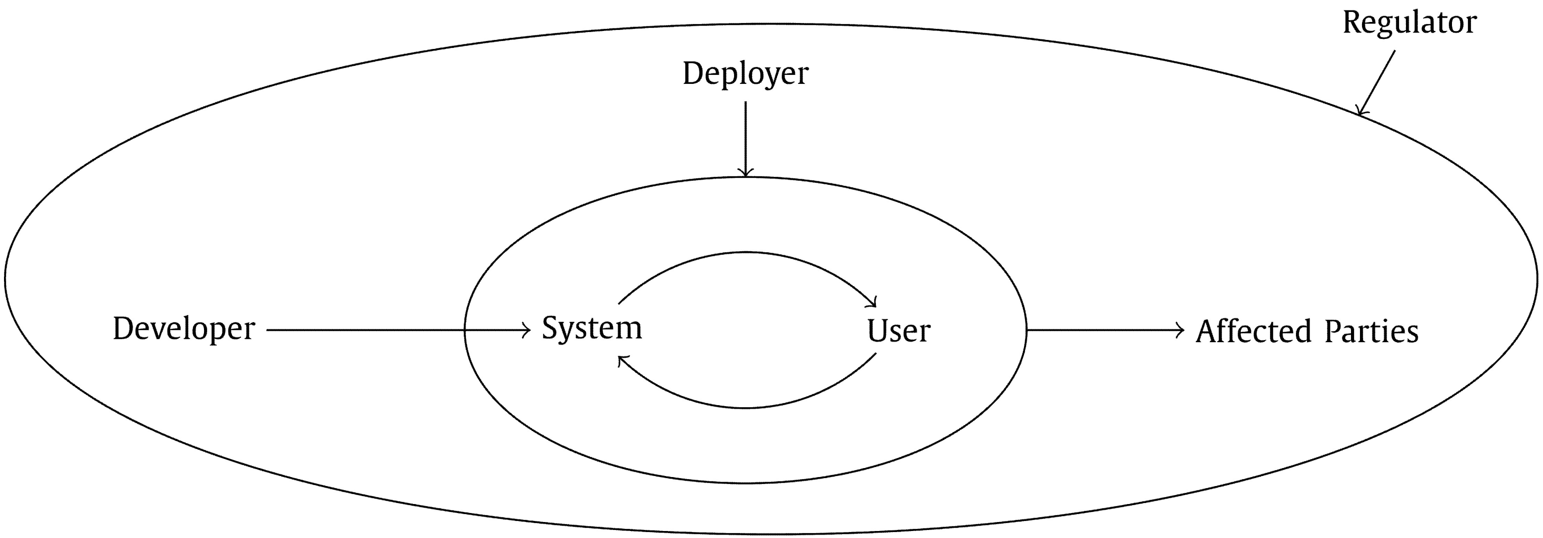

Hello, World:
Artificial intelligence and its
use in the public sector


European
Commission
nce for Policy Report
elligence in public services
and impact
es in the EU
AI
DATA
ROBOTICS
EUR 30255 EN
Joint
Research
Centre

Audiences for explanations in different AI application domains

https://padlet.com/polyxenivassilakopoulou/explanationaudiences

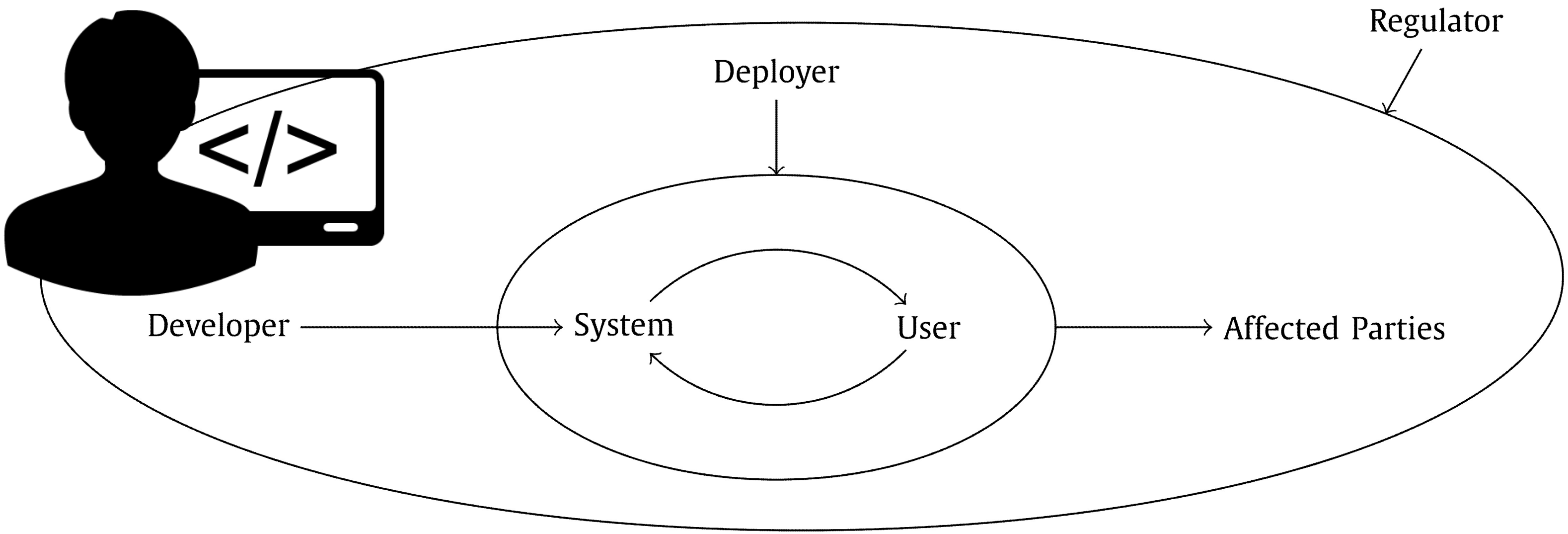
A stakeholder perspective



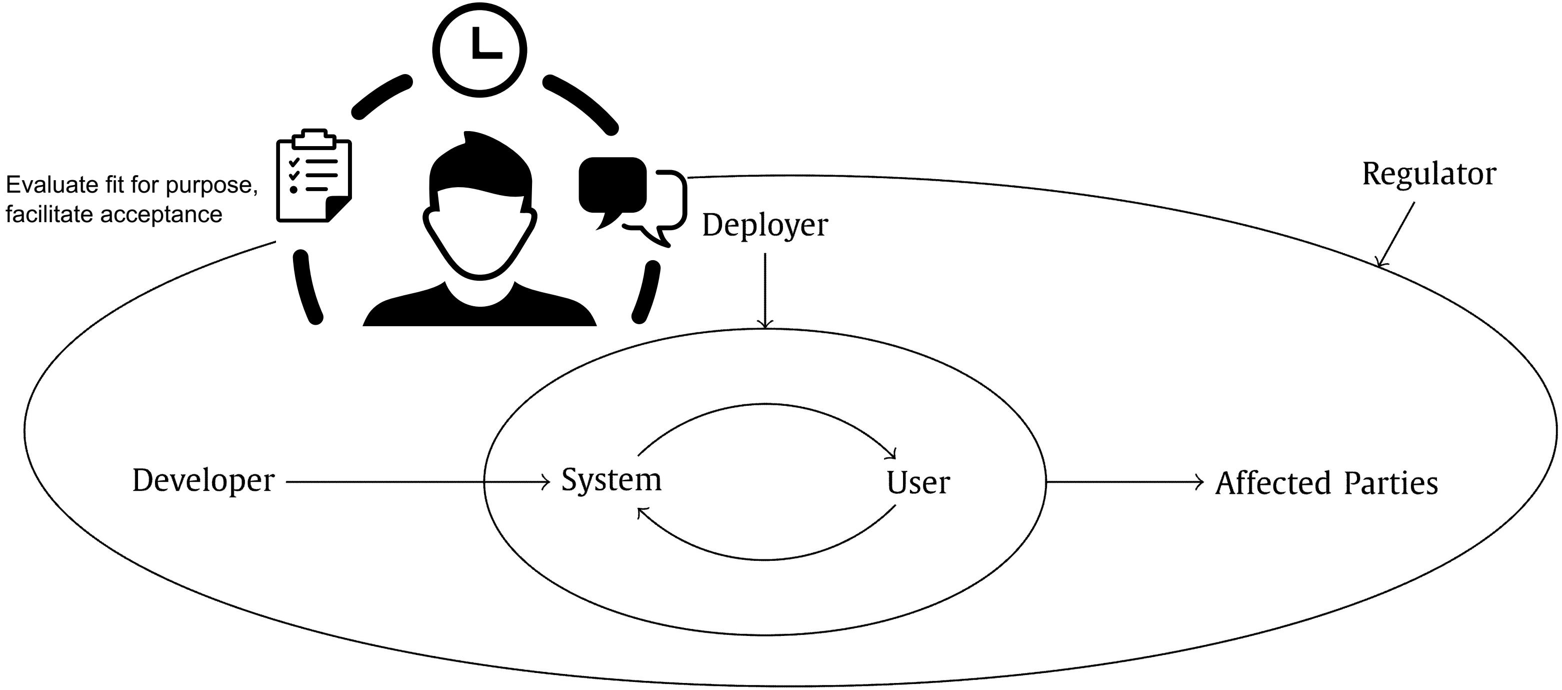
Source: Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence (https://www.sciencedirect.com/science/article/pii/S0004370221000242#)

Different stakeholders have different needs…

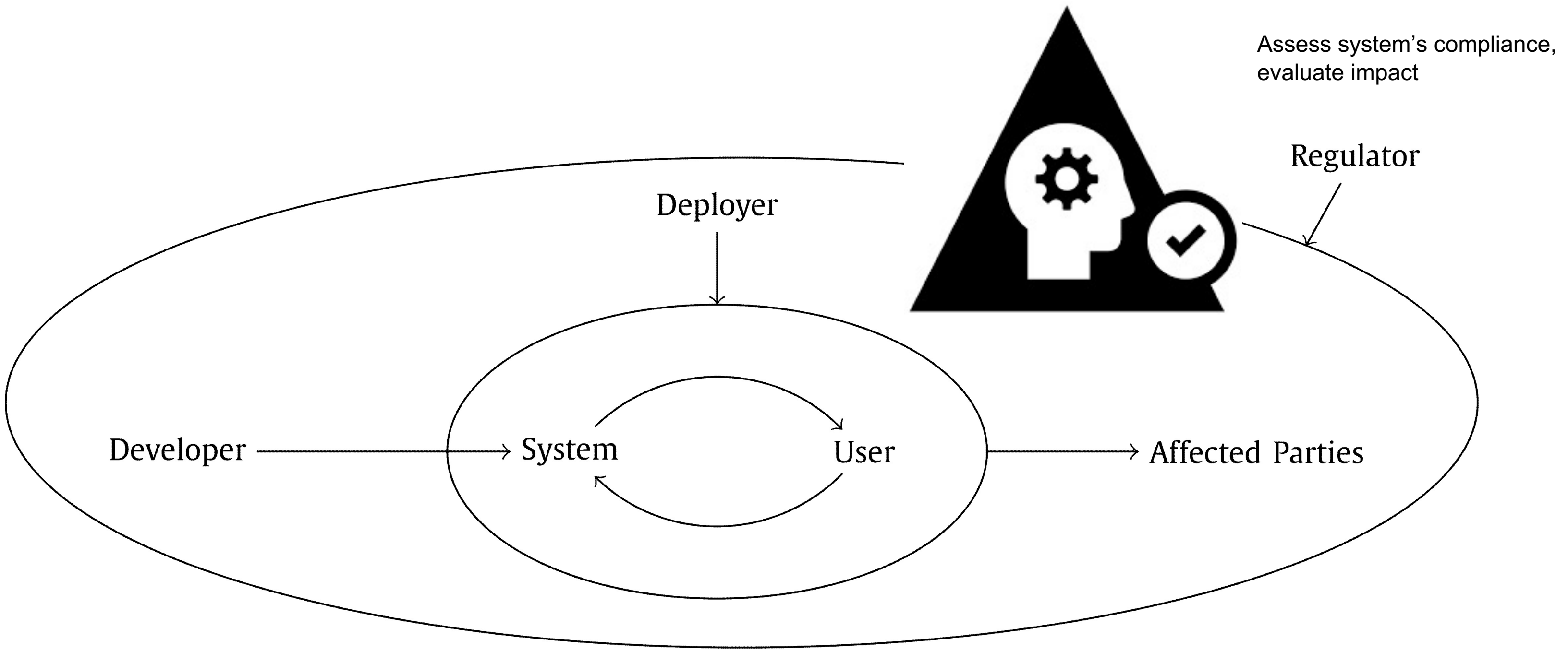
Assess and increase a system's performance e.g. efficiency, predictive accuracy, robustness



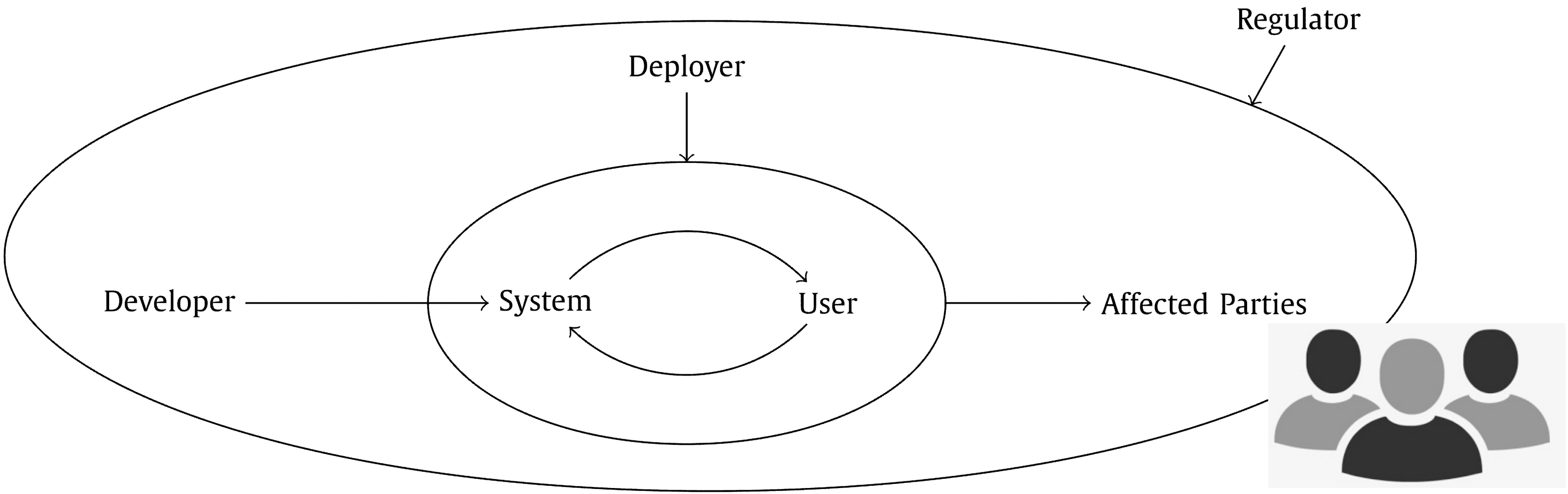
Different stakeholders have different needs…



Different stakeholders have different needs…

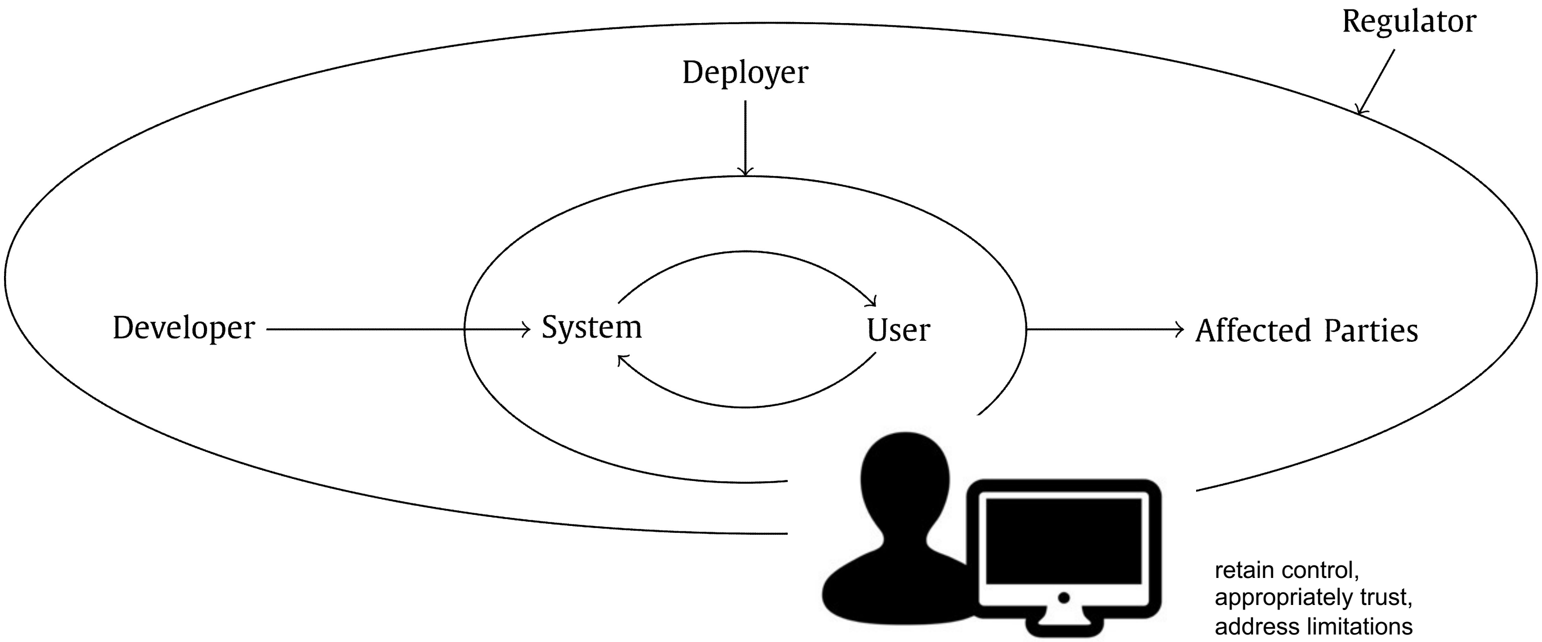


Different stakeholders have different needs…



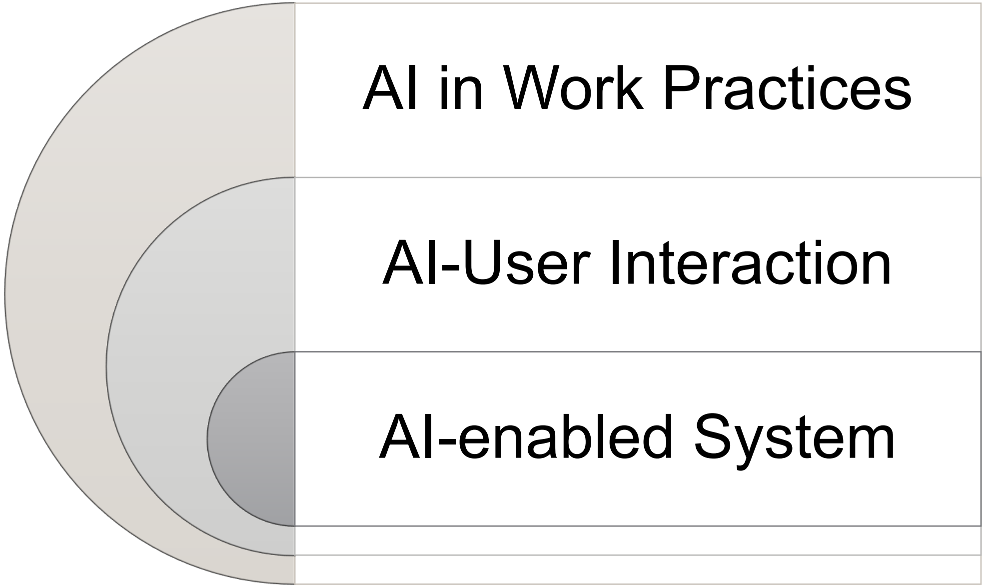
provide informed consent,
challenge outcomes,
insight into personal data processed
right to explanation for profiling

Different stakeholders have different needs...



retain control,
appropriately trust,
address limitations

A multilevel perspective on AI explanations



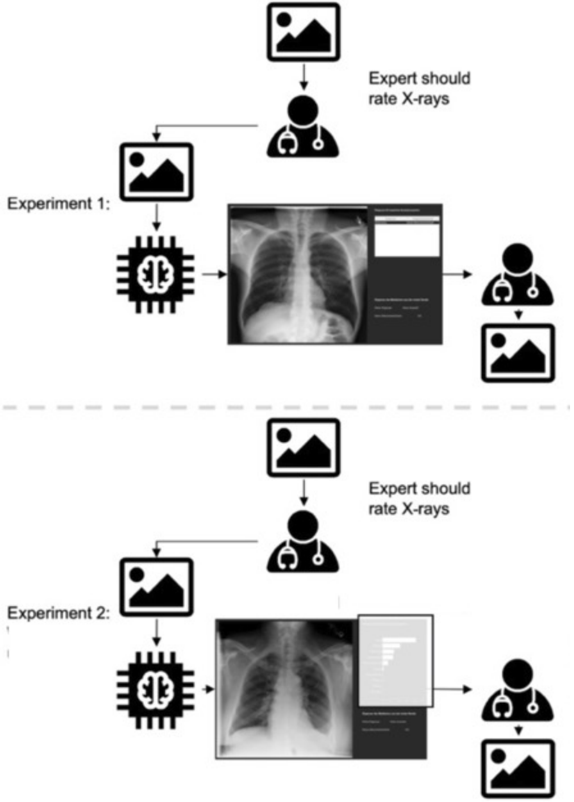
- Studies on the impact of explanations on work performance, human decision making and learning
- Studies on how people relate to different types of explanations e.g how do they rate them in terms of intuitiveness and understandability
- Studies on how well explanations reflect system behaviour

How do explanations affect work practices?

Explanations and indications of model output uncertainty given to radiologists led to a higher overlap between human and machine diagnoses.

Mutual benefits for both human learning and interactive machine learning.

Abel-Karim B, Pfeuffer N, Rohde G, Hinz O (2020) How and what can humans learn from being in the loop?—Invoking contradition learning as measure to make humans smarter. Ger J Artif Intell 34:199–207

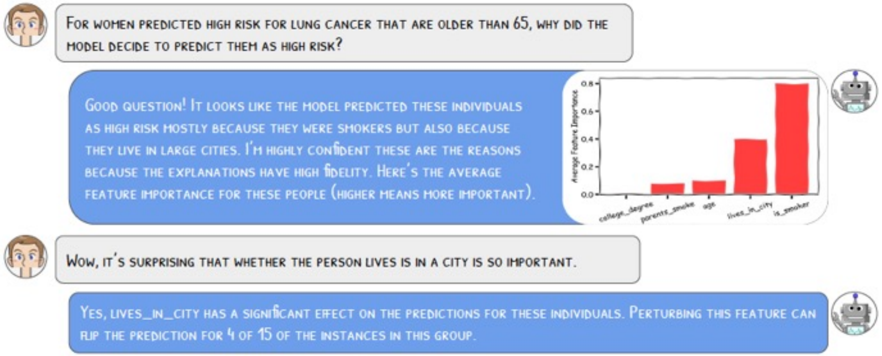


However, other researchers found no evidence that explanations have an effect in trust calibration or even a reduction of trust (BUT: different tasks: income prediction, apartment selling price prediction).

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 295-305.
Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021, May). Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-52).

AI as a co-worker?

Research shows that in knowledge-intensive settings users wish to treat AI as “another colleague”. Existing explanation provision paradigms need to evolve. Users have indicated they would rather enter in dialogue about why a particular recommendation is given (interactive explanations).



Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*

Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). Rethinking Explainability as a Dialogue: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01875*.

UiO : Institutt for informatikk

UiA Universitetet i Agder

AI4Users

Thanks for participating!

Alexander Kempton (University of Oslo) and Polyxeni Vassilakopoulou (University of Agder)