Explainability Auditing for Intelligent Systems:

A Rationale for Multi-Disciplinary Perspectives

Vision Paper

Markus Langer, Kevin Baum, Kathrin Hartmann, Stefan Hessel, Timo Speith, Jonas Wahl

RE4ES: First International Workshop on Requirements Engineering for Explainable Systems

Explainability

- Ability to explain the outputs and internal processes of intelligent systems to relevant stakeholders (e.g. developers, deployer, end users, human in the loop), so that
    - stakeholders can decide whether their interests, needs and demands are met;
    - the system can be improved more easily with respect to a given desideratum.
- Commonly featured concept in regulatory guidelines on AI.

Explainability

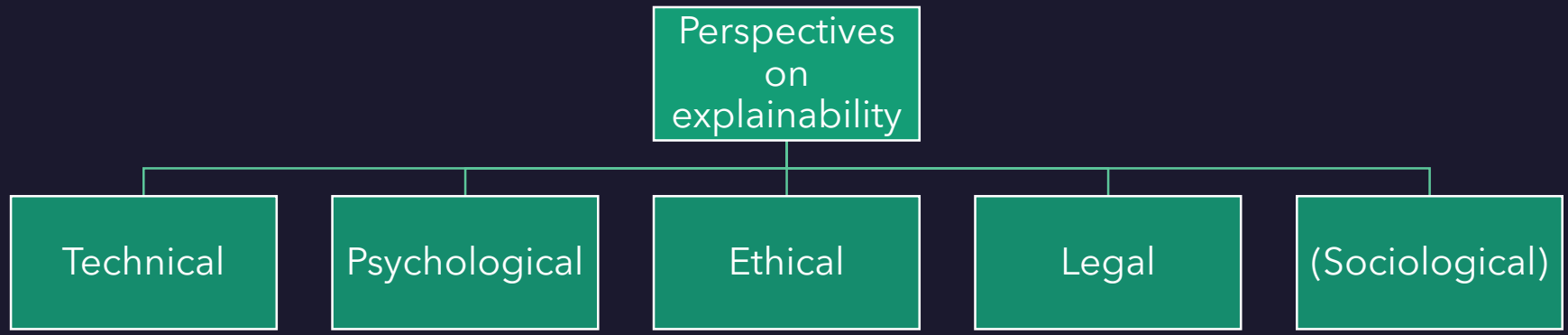
- what is considered a 'good explanation' will depend on the stakeholder to whom the explanation is given.

Ensuring and assessing system explainability is a highly complex task that requires a multi-disciplinary approach.

How to assess explainability?

- Ways to ensure that explainability requirements are fulfilled are
  - explainability audits by specialized auditing firms or institutions,
  - explainability certification, as a means to communicate that a system has undergone quality control.
- Contribution of our vision paper: present a (non-exhaustive) list of explainability requirements based on research in different disciplines.
- See references of our article for more potential requirements.

The different dimensions of explainability



The different dimensions of explainability

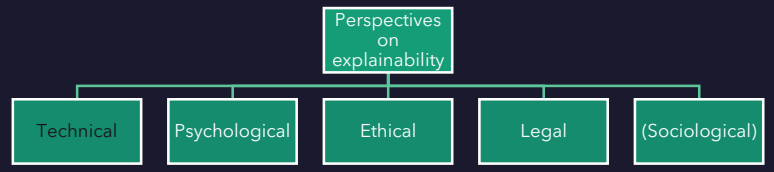
Technical dimensions:

- *Ante-hoc vs. Post-hoc:*

  Is the system designed to allow for human insight or are additional methods employed to generate explanations?

- *Global vs. Local:*

  Is the decision-making process explained as a whole or with regards to specific outputs?



The different dimensions of explainability

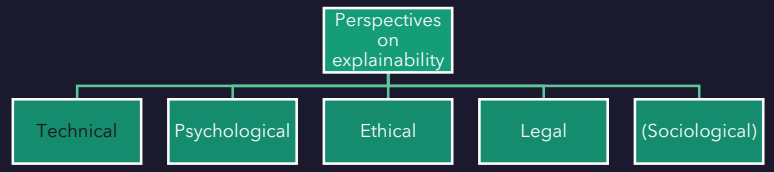
Technical dimensions:

- *Interactive explainability:*

  Can the system's explanations be adapted to the stakeholder's needs or is there a one-size-fits all solution?

- *Explainability trade-off:*

  Is there a trade-off between explainability and accuracy/performance and are choices for one or the other justified?



The different dimensions of explainability

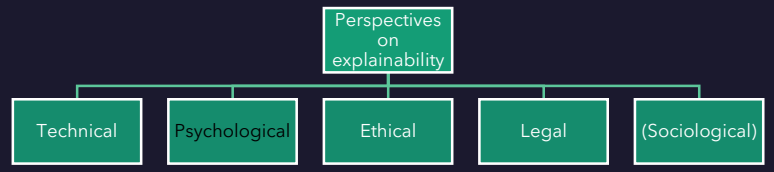
Psychological dimensions:

- *Understandability:*

  Does the system provide explanations that helps people to gain a better understanding of decision processes?

- *Context-Dependency:*

  Does the system provide context-related intellegibility of its decision-making?



The different dimensions of explainability

Psychological dimensions:

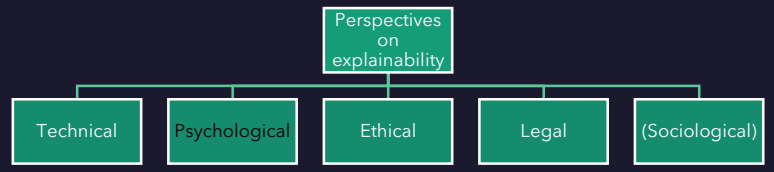
- *Usability:*

  Is the provided information easy to access and easy to use?

- *Honesty:*

  Is the provided information non-deceptive?

  Example: An explanation could be phrased so that contesting a decision appears pointless.

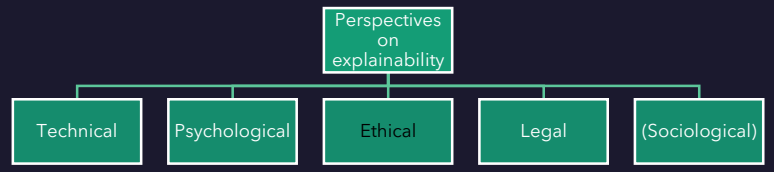


The different dimensions of explainability

Ethical dimensions:

- *Responsibility:*
  - Does the system provide information that enables responsible decision-making (e.g. for a human in the loop)?
  - Does the information make the allocation of moral responsibility possible?

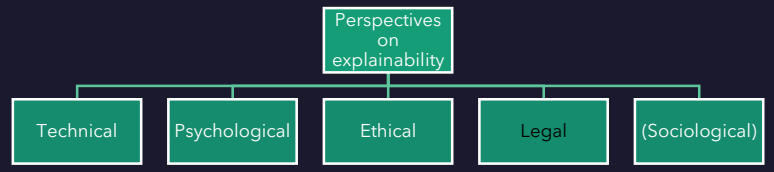
- *Non-Discrimination:*
  - Does the system provide information that makes it possible to detect discrimination?



The different dimensions of explainability

Legal dimensions:

- Does the system comply with
  - *data protection laws*, e.g. GDPR.
  - *cybersecurity laws*, e.g the Cybersecurity Act of the EU.
  - *AI-specific laws and regulations*, e.g. the 'proposal for a regulation laying down harmonized rules on artificial intelligence' of the EU.



Conclusion: the benefits of explainability auditing

- *Technical benefits:*
  - Improved understanding of malfunctions;
  - Improved system safety.
- *Psychological benefits:*
  - Increased system acceptance/trust;
  - Improved human-system performance.
- *Ethical benefits:*
  - Adequate allocation of responsibility;
  - Contesting decisions made easier.
- *Legal benefits:*
  - Compliance with legal obligations can be demonstrated.

Thank you for listening!

More about our work: https://algoright.de/