HCAI-käsitteet

Accountability

Governance and accountability issues refer to those who create the ethics standards for AI, who governs the AI system and data, who maintains the internal controls over the data and who is accountable when unethical practices are identified. (Mintz 2019.)

Accountability can be achieved via complementary means such as algorithmic impact assessments (AIA's), auditing and certification.

Responsibility

Responsible AI means that AI system has all the ethical considerations in place, and it is aligned with the core principles of the developing company. Responsible AI helps organizations to regain control over the AI models that are deployed. (Deloitte 2019.)

Without clear responsibilities, no one is accountable. (Sondergaard 2019.)

Responsibility creates accountability.

Trustworthiness (developers, end-users)

Focus on both communication & interaction

Can be considered as the confidence of whether a model will act as intended when facing a given problem

> Not necessary when "facing a problem" but that the solution perform as expected, is reliable and able to achieve desired goals (?) (SAL)

> What is considered as "problem", e.g. how to spot errors and need for human judgement? (SAL)

The ability to help people understand given system's decisions through e.g. explanations will provide non-experts with a greater sense of trust and will contribute to the users' willingness to continue using AI systems in the future as well (e.g. Riedl, 2019).

Trustworthy AI systems need to include policies that clearly establish who is
responsible and accountable
(ethical considerations) for their output. An organizational structure and policies should be put in place that can help clearly determine who is responsible for the output of AI system
decisions. (Saif & Ammanath, 2020.)

Transparency (developers)

Focus on communication;

Creates explainability -> understandability

Model is considered to be transparent if it is by itself understandable (refers to the characteristic of a model to be, on its own, understandable for a human (Arrieta et al. 2020). The opposite of 'black-box-ness'.

> Even though a model is transparent, it might not be understandable for an end-user. If you as a designer don't know how AI system makes decision, you cannot explain it. (SAL)

Transparency should not be the ultimate solution for users or people affected by the algorithmic decisions since source code is illegible to non-experts.

Requires technical steps and technical correctness: the developer of the model must be able to explain

Means using intuitive language to talk about the AI systems under a development, how they work and what they are capable of (MindAI 2018)

Requires consideration whether the outcomes of the model are statistically sound: whether certain groups are under-represented in the outcomes (related to fairness)

Privacy awareness (developers, end-users)

Focus on communication (related to transparency?)

ML model's ability to assess privacy. ML models can have complex representations of their learned patterns. Not being able to understand what the model has captured and stored can entail a privacy breach (e.g. Castelvecchi,2016).

Explainability (end-users)

Focus on interaction;
Creates understanding

Helps users to understand the algorithms and parameters a system uses by, for example, giving a reason for a particular action. (Xu, 2019).

Explainability can be a considered as a bridge to avoid unfair and unethical use of algorithmic outputs (Arrieta et al., 2020).

From a social viewpoint, explainability can be seen as the capacity to reach and guarantee fairness in ML models.

Helps users to understand the systems decision process and helps to create realistic mental models of the system's capabilities and limits (Google People + AI Guidebook 2019, IBM Everyday Ethics for AI 2019).

Optimizing for understanding is recommended instead of completely explaining the system. (Google 2019.)

For any given automated behavior, developers must be able to explain why the algorithm engaged in
that particular course of action. (Ramamoorthy 2019).

Traceability of model predictions; how it arrives at distinct predictions and how it processes the input (Otto / ethics guideliens)

Informativeness (end-users)

Focus on interaction (related to explainability?)

Explainable ML models should give out information about the (perceived) problem they are tackling, i.e. extracting information about the inner relationships of the model (simple proxies) (Arrieta et al., 2020).

Interactivity (end-users)

Focus on interaction (related to explainability?)

Can include the ability of an explainable ML model to be interactive with the user (e.g. Langley, P. et al., 2017).

Interpretability (developers, end-users)

Focus on interaction (related to explainability?)

The ability to explain or to provide the meaning in understandable terms to a human. (e.g. Arrieta et al., 2020)

Intelligibility of a model or system; transparency with regard to how it works and what it does (Otto / ethics guidelines)

Comprehensibility (developers, end-users)

Focus on interaction (related to explainability?)

The ability of a learning algorithm to represent what it has learned/ what it knows in a way that humans can understand.(e.g. Guidotti et al., 2018)

Connected to understandability in that it relies on the capability of the audience to understand the knowledge contained in the model.

Fairness (AI developers & designers)

The requirement that all users are treated equally and without prejudice (e.g. Riedl, 2019).

From a social viewpoint, explainability can be seen as the capacity to reach and guarantee fairness in ML models. (e.g. Riedl, 2019).

Fairness is related to trustworthy AI: AI must be designed and trained to follow a fair, consistent process and make fair decisions. To avoid problems related
to fairness and bias, companies first need to determine what constitutes "fair."(Saif & Ammanath 2020)

Another way of 'screening for fairness' is highlighting the bias in the data the model was exposed to (e.g. Hendricks et al., 2017).

An explainable ML model can be e.g. a clear visualization of the reasons affecting the result, allowing the verification of fairness or ethical evaluation of the model (e.g. Goodman, B., & Flaxman, S., 2017).