# UNPACKING THE PURPOSES OF EXPLAINABLE AI

**Suzana Alpsancar, Tobias Matzner, Martina Philippi**

Paderborn University (Germany)

suzana.alpsancar@uni-paderborn.de; tobias.matzner@uni-paderborn.de,
martina.philippi@uni-paderborn.de

## EXTENDED ABSTRACT

AI systems are being increasingly deployed in various societal fields. Much of the AI technology that is contributing to this success, particularly machine learning (ML), is opaque, meaning that how it works or why it exhibits a particular behavior or performance is not (immediately) obvious for a variety of reasons (Burrell 2016). Numerous cases have shown that this opacity can become problematic: Some ML models have been easy to trick and have exhibited Clever-Hans effects, domain shifts, and overfitting. Others have incorrectly influenced grave decisions such as the probability of death in a patient with pneumonia (Cabitza, Rasoini, and Gensini 2017), or have been subject to adversarial attacks (Gilpin et al. 2018). Scandals and debates about the biases and fairness of, for example, COMPAS recidivism prediction software (Angwin et al. 2016) have contributed to growing ethical concerns about AI. According to the literature review by Tsamados et al. (2022), these ethical concerns can be distinguished into two normative concerns (unfair outcomes and transformative effects), three epistemic concerns (inconclusive evidence, inscrutable evidence, and misguided evidence), as well as the concern of traceability (the possibility of tracing the chain of events of factors that brought about a given outcome) that affects all other concerns. Whereas the normative concerns relate explicitly to ethical impacts such as unintended consequences or biases of AI systems, the epistemic concerns relate to the justifiability of the outcome of AI systems, and this, in turn, may evoke morally critical decisions.

For all these reasons, it would be game changing if both adopters (e.g., medical practitioners) and affected individuals (e.g., patients) would be able to adequately assess the performance and limitations of AI systems. There is a widespread at least implicit assumption in the field that ”explainability is a suitable means for facilitating trust in a stakeholder”, what Kästner et al. (2021) have depicted as the ”Explainability-Trust-Hypotheses”. Against this background, explainable AI (xAI) has become highly valorized. Explainability is considered as necessary for robust and trustworthy AI applications and, hence, for their commercial success (Arya et al. 2019). As several meta-reviews have shown (Hagendorff 2020; Jobin, Ienca, and Vayena 2019; Morley et al. 2020), explainability is a central element of all voluntary commitments and ethical guidelines for AI in industry, research, and policymaking. For instance, the European Commission’s High Level Expert Group on Artificial Intelligence literally links the research field of xAI to its agenda of building trustworthy AI:

> For a system to be trustworthy, we must be able to understand why it behaved a certain way and why it provided a given interpretation. A whole field of research, Explainable AI (xAI) tries to address this issue to better understand the system’s underlying mechanism and final solutions. (High Level Expert Group 2019, 21)

The latest regulatory requirements echo this valorization of explainability, especially within the EU where legislation might expand existing laws into a right to explanation (Wachter, Mittelstadt, and Floridi 2017) and where the proposed AI Act sets new obligations to ensure transparency, which is often directly linked to explainability (EC 2021).

From a philosophical perspective, the call for xAI rests on a normative claim: "good AI is xAI" or even the stronger claim "only xAI is good AI." This valorization runs the risk of being overgeneralized because explanations are not per se useful, appropriate, or demanded. Clearly, the practical use of xAI depends on whether the explanation is needed at all, whether it is appropriate for the explainees, and whether it is understandable. Previous literature reveals some voices that are critical of the value of explaining. For instance, Robbins argues that the principle of explicability[1] is misdirected. He points out three misgivings: (1) It is not the process of coming up with a decision, but the decision (or action) itself that is in need of an explanation. (2) It makes no sense to demand from all AI systems that they should explain themselves, because there are many applications with a low risk (in terms of potential harm of moral weight). (3) For high-risk applications, it is contradictory to demand explicability from the AI system, because they are designed precisely to serve areas in which we do not know what parameters to consider (Robbins 2019, 509).

We agree with Robbins' basic intervention that not all AI systems must necessarily be explainable, that explainability is not a value in itself, and that explainability is not always useful. However, we disagree with his classificatory theoretical perspective: Neither algorithms nor decisions can be classified per se as needing or not needing explanations—which is what he suggests as being a better strategy. Instead, we follow a practice theoretical approach in arguing that explainability should neither be conceptualized as a trait of a technical artifact nor as a property of a mere decision or an act, but as a disposition of a given sociotechnical system that must be materialized in practices of explaining within given socially structured contexts.

If we account for explainability as an instrumental value, we need to explicate what explainability is meant to deliver from both an ethical perspective and the perspective of respective users (or other stakeholders). Hence, we need to answer the following normative questions when it comes to adequately evaluating the goodness of explanations:

1. When is an explanation ethically obligatory?
2. When is an explanation individually helpful (to whom for which purpose)?
3. What characterizes a good explanation (in light of 1. and 2.)?

Currently, these rarely explicated questions are usually answered by referring to those motives that give reasons for developing xAI in the first place—that is, naming what xAI is meant to be good for. These for-the-sake-of relations can be systematized into three categories:

a. Functional purposes such as keeping a system running, debugging it, or improving it technically (developing AI)

[1] Robbins adopts the language of Floridi et al. (Floridi et al. 2018) and argues against the claim that all AI must be explicable in their sense, that is of guaranteeing "meaningful human control." His objections, however, can be related to a generalization of explainability, not just to its utility for this interpretation of "ethical assurance."

b. Social or economic purposes such as satisfying so-called users' needs, e.g., explaining apparently awkward social robot behavior (deploying AI)

c. Normative purposes, i.e. respecting ethical values and principles or meeting legal requirements, e.g., presenting reasons for loan rejections to render the decision-making process contestable (governing AI)

The first type of purposes echoes the experiences of those who develop and optimize ML components, e.g., the first techniques for explaining AI had been developed by ML experts for other experts, e.g., in the context of the Pascal Visual Object Classes (VOC) Challenge, which serves as a benchmark for object recognition/detection in ML, to unmask Clever-Hans effects (Everingham et al. 2015). With the wider distribution of AI systems (AIS) in various societal fields, the xAI community increasingly draws attention to lay persons (users, operators, domain experts) and to meet ethical and legal demands. Here, there is a strong motivation to mimic interpersonal interaction. For instance, de Graaf and Malle (2017) argue that the entire interaction with nonhuman agents, including explanations, should correspond to the user's expectations, namely their underlying intentional framework. If AI systems do not reveal their intention, users find them "unsettling and creepy" (de Graaf and Malle 2017, 20).

In terms of the ethical demands, much has been said about the challenges of moving 'from principles to practice' (Rességuier and Rodrigues 2020). Very little has been discussed about the conditions under which certain purposes can be considered adequate: When is it necessary, helpful, or adequate for an xAI system to serve the purpose of particular ethical principles, and how does this relate to other purposes xAI is meant to serve?

In our paper we aim to put the goodness of the presumed purposes, xAI is meant to serve, into question and we want to particularly question how functional, economic, and ethical purposes relate to each other. As we think that such an evaluation only makes sense in a contextualized setting, we will pursue our analyses by comparing two stylized use cases: deploying automated and connected vehicles and deploying algorithmic decision-making systems in a healthcare facility.

**KEYWORDS:** Explainable AI, valorization of xAI, purposes, ethical demands, users' needs.

## REFERENCES

Angwin, Julia, Larson, Jeff, Mattu, Surya and Kirchner, Lauren Kirchner (2016). "Machine Bias. There's Software used across the country to predict future criminals. And it's biased against blacks." ProPublica, accessed March 27, 2022. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arya, Vijay, Bellamy, Rachel K. E., Chen, Pin-Yu, Dhurandhar, Amit, Hind, Michael, Hoffmann, Samuel C., Houde, Stephanie, et al (2019). "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques." CoRR abs/1909.03012. arXiv: 1909.03012

Burrell, Jenna (2016). "How the machine thinks: Understanding opacity in machine learning algorithms." Big Data & Society 3 (1): 1–12. https://doi. org/10.1177/2053951715622512

Cabitza, Federico, Rasoini, Raffaele, and Gensini, Gian Franco (2017). “Unintended Consequences of Machine Learning in Medicine.” JAMA 318 (6): 517–518. https://doi.org/10.1001/jama.2017.7797

De Graaf, Maartje MA, and Malle, Bertram F. (2017). “How people explain action (and autonomous intelligent systems should too).” In 2017 AAAI Fall Symposium Series.

EC (2021). “Proposal for regulation of the European parliament and of the council Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts.” European Commission. Digital Strategy, accessed September 12, 2022. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

Everingham, Mark, Eslami, SM Ali, Van Gool, Luc, Williams, Christopher KI, Winn, John and Zisserman Andrew (2015). “The pascal visual object classes challenge: A retrospective.” International journal of computer vision 111: 98–136. https://doi.org/10.1007/s11263-014-0733-5.

Floridi, Luciano, Cowls, Josh, Beltrametti, Monica, Chatila, Raja, Chazerand, Patrice, Dignum, Virginia, Luetge, Christoph et al (2018). “AI4People–An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.” Minds and Machines 28 (4): 689–707. https://doi. org/10.1007/s11023-018-9482-5.

Gilpin, Leilani H, Bau, David, Yuan, Ben Z, Bajwa, Ayesha, Specter, Michael and Kagal, Lalana (2018). “Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning.” In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics.

Hagendorff, Thilo (2020). “The Ethics of AI Ethics: An Evaluation of Guidelines.” Minds and Machines 30 (1): 99–120. https://doi.org/10.1007/s11023-020-09517-8

High Level Expert Group (2019). “Ethics Guidelines for Trustworthy AI.” European Commission, accessed October 29, 2021. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Jobin, Anna, Ienca, Marcello and Vayena, Effy (2019). “Artificial Intelligence: the global landscape of ethics guidelines.” Nat. Mach. Intell., 389–399. https: //doi.org/10.1038/s42256-019-0088-2

Kästner, Lena, Langer, Markus, Lazar, Veronika, Schomäcker, Astrid, Speith, Timo and Sterz, Sarah (2021). “On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness.” In 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), 169–175. https://doi.org/10.1109/REW53955.2021.00031

Morley, Jessica, Floridi, Luciano, Kinsey, Libby and Elhalal, Anat (2020). “From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices.” Science and Engineering Ethics 26:2141–2168. https://doi.org/10.2139/ssrn.3830348

Rességuier, Anais, and Rodrigues, Rowena (2020). “AI ethics should not remain toothless! A call to Bring back the teeth of ethics.” Big Data & Society 7 (2): 1–5. https://doi.org/10.1177/2053951720942541

Robbins, Scott (2019). “A Misdirected Principle with a Catch: Explicability for AI.” Minds and Machines 29 (4): 495–514. https://doi.org/10.1007/s11023-019-09509-3

Tsamados, Andreas, Aggarwal, Nikita, Cowls, Josh, Morley, Jessica, Roberts, Huw, Taddeo, Mariarosaria and Floridi, Luciano (2022). “The ethics of algorithms: key problems and solutions.” AI & SOCIETY 37 (1): 215–230. https://doi.org/10.1007/s00146-021-01154-8

Wachter, Sandra, Mittelstadt, Brent and Floridi, Luciano (2017). “Why a right to explanation of automated decision-making does not exist in the general data protection regulation.” International Data Privacy Law 7 (2): 76–99.