

trustworthy artificial intelligence (trustworthy AI)

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)

Editors: Konstantina Oliumtsevs (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

Trustworthy artificial intelligence (trustworthy AI) names an artificial intelligence system (AI system) that meets requirements beyond accuracy: human oversight, robustness, privacy, transparency, fairness, societal well-being, and accountability. The seven key requirements of the EU Ethics Guidelines are stated for the AI system as a whole, including the people and the organization around the machine learning (ML) method inside it. Five of them constrain the components of empirical risk minimization (ERM): the training set, the model, the loss function, the algorithm, and the learned hypothesis. Each such requirement is a measurable property of that hypothesis or of the method that learned it: a bound on the change of a prediction under a perturbation of the feature vector, the change of the prediction when only a sensitive attribute is flipped, the counterfactual that explains a rejection, or the share of borderline decisions passed to a person. The EU AI Act makes several of these requirements binding for a high-risk artificial intelligence system (high-risk AI system), and they can conflict, so trustworthiness is a set of measured properties balanced for the application rather than a single number.

Definition

Consider a bank that uses an artificial intelligence system (AI system) to decide on loan applications. The system applies a hypothesis $\hat{h}$, learned from the records of completed loans, to the feature vector $\mathbf{x}$ of an applicant, which collects the income, the gender, and further attributes. The loan is approved when the predicted creditworthiness $\hat{h}(\mathbf{x})$ exceeds a threshold τ . The bank, its supervisor, and the applicant rely on these decisions only if the system meets requirements beyond accuracy. A small change of the income must not flip the decision, the gender alone must not flip it either, a rejected applicant must learn why, and a loan officer must be able to overrule the system. Trustworthy AI names an AI system that meets such requirements.

The Ethics Guidelines of the European Commission’s High-Level Expert Group state seven key requirements (KRs) (High-Level Expert Group on Artificial Intelligence, 2019): human agency and oversight (KR1), technical robustness and safety (KR2), privacy and data governance (KR3), transparency

(KR4), diversity, nondiscrimination, and fairness (KR5), societal and environmental well-being (KR6), and accountability (KR7). The accompanying assessment list turns each requirement into questions that the provider and the deployer of the AI system answer about their system (High-Level Expert Group on Artificial Intelligence, 2020).

The requirements are stated for an AI system as a whole, which includes the data pipeline, the user interface, and the organization that operates it, and not only for the machine learning (ML) method inside it. KR6 and KR7 concern the purpose and the governance of the system and have no counterpart in the ML method. KR1 to KR5, in contrast, constrain the components of empirical risk minimization (ERM): the training set $\mathcal{D}^{(\text{train})}$, the model $\mathcal{H}$, the loss function, and an algorithm that minimizes the average loss over $\mathcal{D}^{(\text{train})}$, as well as the learned hypothesis $\hat{h}$ that results (Pfau and Jung, 2024). Each of these requirements can be stated as a measurable property of $\hat{h}$ or of the method that learned it, and the design choices for the components of ERM are the means to meet it.

Fig. 1 places KR1, KR2, KR4, and KR5 in the credit-scoring example. The horizontal axis is a one-dimensional sketch of the feature vector $\mathbf{x}$: the income, the gender, and the further attributes of an applicant collapsed into one coordinate. The vertical axis is the creditworthiness, the label y . The filled circles are the training set $\mathcal{D}^{(\text{train})}$ of $m = 5$ completed loans, the solid curve the hypothesis $\hat{h}$ learned by minimizing the average loss over $\mathcal{D}^{(\text{train})}$, and the dotted line the approval threshold $\tau = 3.8$. The applicant with feature vector $\mathbf{x}'$ (open square) obtains the predicted creditworthiness $\hat{h}(\mathbf{x}') = 2.94$ and is rejected.

KR2 asks that a small perturbation of the feature vector changes the prediction little (Pfau and Jung, 2024, Sect. V). The perturbation of $\mathbf{x}'$ by $\delta = 0.5$ (horizontal arrow) moves the predicted creditworthiness from 2.94 to 3.19 (vertical arrow). The slope of $\hat{h}$ never exceeds 1.225, so no perturbation by δ changes the prediction by more than 1.225δ : the learned hypothesis is Lipschitz continuous with constant 1.225 (see robustness). KR2 also covers perturbations of $\mathcal{D}^{(\text{train})}$, e.g., by data poisoning, against which the stability of the ML method protects.

KR5 asks that the decision does not depend on a sensitive attribute s such as gender. The open triangle depicts the same applicant with only the gender flipped, the feature vector $\mathbf{x}^*$. The predicted creditworthiness rises to $\hat{h}(\mathbf{x}^*) = 3.93$, above τ , and the loan is approved: the decision depends on the gender, which violates the counterfactual criterion of fairness. The gap $\hat{h}(\mathbf{x}^*) - \hat{h}(\mathbf{x}') = 0.99$ measures the violation. It is closed by a design choice for a component of ERM: the model (excluding the sensitive attribute and the features correlated with it), the loss function (a penalty term for the gap), or $\mathcal{D}^{(\text{train})}$ (resampling the groups) (Pfau and Jung, 2024, Sect. VIII).

KR4 asks for transparency: the rejected applicant is told that an AI system was used and may obtain an explanation. The counterfactual $\mathbf{x}''$ (open diamond), at which $\hat{h}$ reaches τ , is such an explanation. It is the smallest change of the income, at fixed gender, that flips the decision, and it tells the applicant

the income at which the loan would be approved. The gender flip $\mathbf{x}^*$ also flips the decision, but it is not a change the applicant can make, which is what makes it a violation of KR5 rather than an explanation.

KR1 asks that a person can oversee and overrule the decisions. The shaded band marks predicted creditworthiness within $\varepsilon = 0.3$ of τ . Applications in this band are passed to a loan officer, who decides; one of the five loans in $\mathcal{D}^{(\text{train})}$ falls into it. A wider band passes more applications to the officer and leaves fewer borderline decisions to the system (see human oversight). KR3 asks that the records of completed loans, which are personal data, are protected during training and cannot be recovered from $\hat{h}$; differential privacy (DP) and federated learning (FL) are design choices for the algorithm and for $\mathcal{D}^{(\text{train})}$ to that end (see data governance).

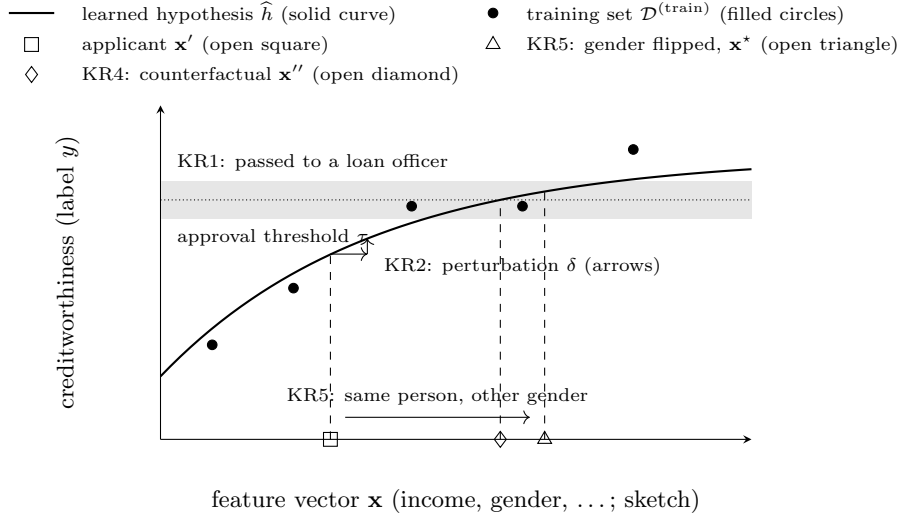


Figure 1: Four key requirements for trustworthy AI in the credit-scoring example. The horizontal axis is a one-dimensional sketch of the feature vector. The solid curve is the hypothesis $\hat{h}$ learned from the training set $\mathcal{D}^{(\text{train})}$ of five completed loans (filled circles); a loan is approved when the predicted creditworthiness exceeds the threshold τ (dotted line). The applicant with feature vector $\mathbf{x}'$ (open square) is rejected. KR1: predictions in the shaded band are passed to a loan officer. KR2: a perturbation of $\mathbf{x}'$ by δ (horizontal arrow) changes the prediction little (vertical arrow). KR4: the counterfactual $\mathbf{x}''$ (open diamond), the smallest change of the feature vector at which $\hat{h}$ reaches τ , explains the rejection. KR5: the same applicant with the gender flipped, $\mathbf{x}^*$ (open triangle), is approved, so the decision depends on the gender

The requirements are increasingly codified. The EU AI Act makes KR1 to KR4 binding for a high-risk artificial intelligence system (high-risk AI system), a class that includes the credit scoring of natural persons (European Parlia-

ment and Council of the European Union, 2024, Annex III): risk management (Art. 9), data governance (Art. 10), transparency (Art. 13), human oversight (Art. 14), and accuracy, robustness, and cybersecurity (Art. 15) (European Parliament and Council of the European Union, 2024). ISO/IEC 42001 specifies a certifiable management system for artificial intelligence (AI) (International Organization for Standardization and International Electrotechnical Commission, 2023), and the AI Risk Management Framework of the US National Institute of Standards and Technology structures trustworthiness into the functions govern, map, measure, and manage (National Institute of Standards and Technology, 2023). The requirements can conflict: closing the gender gap raises the average loss on $\mathcal{D}^{(\text{train})}$, and the noise that DP adds to the training makes explanations less precise (Pfau and Jung, 2024). Trustworthiness is therefore a set of measured properties that are balanced for the application, not a single number.

Cross References

AI system, EU AI Act, high-risk AI system, provider, deployer, human oversight, robustness, stability, data poisoning, data governance, DP, transparency, explainability, counterfactual, fairness, sensitive attribute, ERM

References

1. High-Level Expert Group on Artificial Intelligence (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI): For self assessment. European Commission. URL: <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>.
2. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.
3. High-Level Expert Group on Artificial Intelligence (2019). Ethics Guidelines for Trustworthy AI. European Commission. URL: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.
4. International Organization for Standardization and International Electrotechnical Commission (2023). ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system. International Standard, 1st edition. URL: <https://www.iso.org/standard/42001>.

5. National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. URL: <https://doi.org/10.6028/NIST.AI.100-1>.
6. Pfau and Jung (2024). Engineering Trustworthy AI: A Developer Guide for Empirical Risk Minimization. arXiv preprint arXiv:2410.19361. URL: <https://arxiv.org/abs/2410.19361>.