

Operationalising AI bias standards

ADAM LEON SMITH FBCS

Technical Non-Executive Director

Chair, BCS Fellows Technical Advisory Group

Board member, ForHumanity

ISO/IEC and CEN/CENELEC AI standardisation delegate

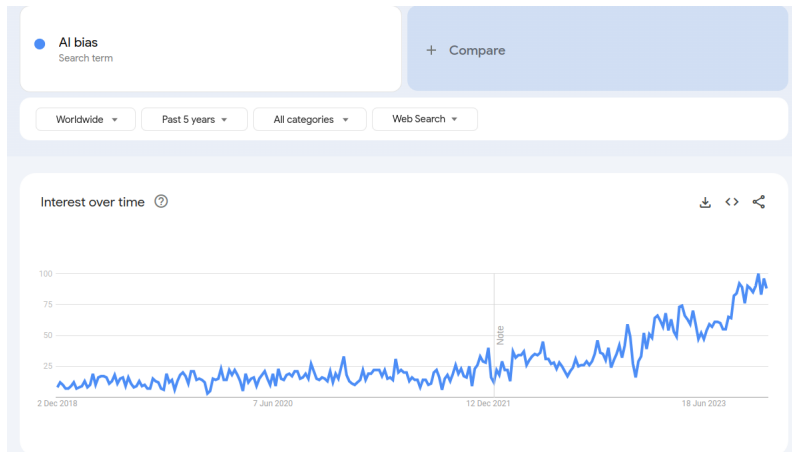
4th December 2023

What is bias?

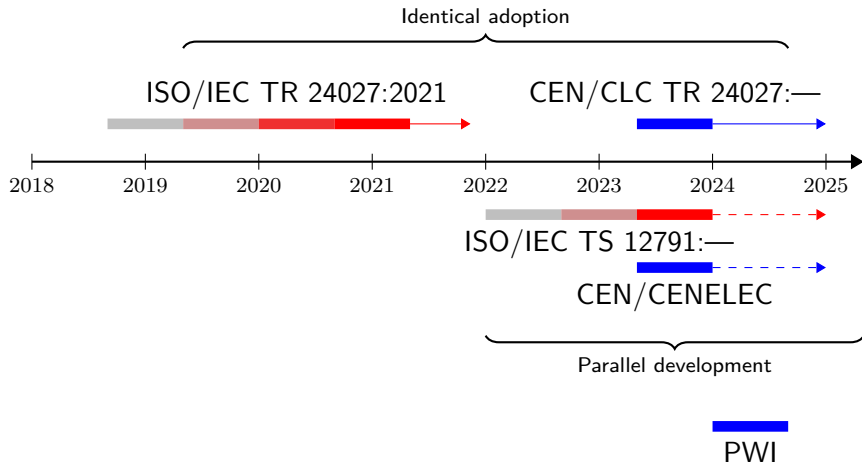
bias (n.)

- ▶ 1520s, *oblique or diagonal line*, from French *biais a slant, a slope, an oblique*, also figuratively, *an expedient, means* (13c., originally in Old French a past-participle adjective, *sideways, askance, against the grain*, a word of unknown origin.
- ▶ Probably it came to French from Old Provençal *biais*, which has cognates in Old Catalan and Sardinian, and is possibly via Vulgar Latin.
- ▶ In the old game of bowls, it was a technical term used in reference to balls made with a greater weight on one side, causing them to curve obliquely (1560s); hence the figurative use *a one-sided tendency of the mind* (1570s), and, at first especially in law, *undue propensity or prejudice*.

Concerns about bias and AI



Bias and AI: Standards activity



What does TR 24027 say?

ISO/IEC TR 24027:2021 — Information technology — Artificial intelligence (AI) — **Bias in AI systems and AI aided decision making**

This document addresses bias in relation to AI systems, especially with regards to AI-aided decision-making. Measurement techniques and methods for assessing bias are described, with the aim to address and treat bias-related vulnerabilities. All AI system lifecycle phases are in scope, including but not limited to data collection, training, continual learning, design, testing, evaluation and use.

What does TR 24027 say?

The Technical Report:

- ▶ Defines terms!
- ▶ Positions unwanted bias in the context of fairness
- ▶ Defines high-level groups of bias sources
- ▶ Defines 23 types of bias
- ▶ Describes assessment processes
- ▶ Describes possible treatments

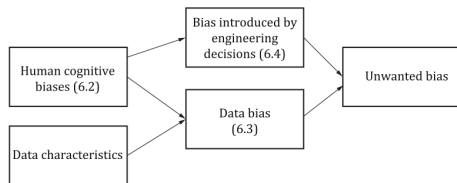


Fig 1: Relationship between high-level groups of bias.
ISO/IEC TR 24027, 6.1

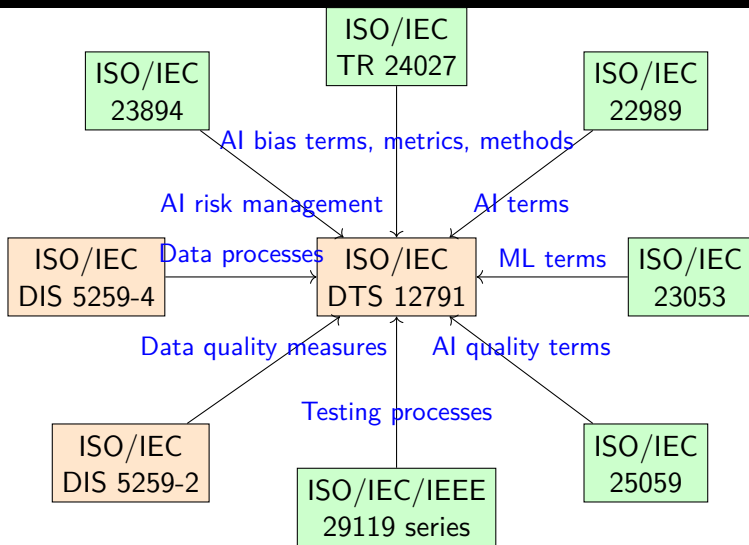


Figure: Fig 2: Relationship to other standards

Key terms

bias systematic difference in treatment of certain objects, people, or groups in comparison to others

at-risk group subset of stakeholders that can be adversely affected by unwanted bias

functional correctness degree to which a product or system provides the correct results with the needed degree of precision

representative training data training data where the size, distribution, and composition are sufficient to train an ML model that achieves an acceptable level of functional correctness for the target population.

What does ISO/IEC DTS 12791 say?

ISO/IEC DTS 12791 — Information technology — Artificial intelligence — **Treatment of unwanted bias in classification and regression machine learning tasks**

This document provides mitigation techniques that can be applied throughout the AI system life cycle in order to treat unwanted bias. This document describes how to address unwanted bias in AI systems that use machine learning to conduct classification and regression tasks. This document is applicable to all types and sizes of organization.

Thank you

email: adamleonsmith@protonmail.com