Towards a Conceptual Framework for Fairness of AI Systems

There is a broad consensus that fairness is crucial to the trustworthiness of AI systems in various critical application scenarios. However, fairness is, by its very nature, a vague concept that is subject to ongoing philosophical, political and social debate, and therefore hard to standardize.
We argue that there is no need for a universal definition of algorithmic fairness, but rather a justifiable case-specific operationalisation.

Operationalizing the concept of fairness requires a clear distinction between the fairness of an algorithm and the fairness of an algorithmic decision process. While not generally sufficient, it is necessary for an algorithmic decision process to be fair that the algorithm used sufficiently satisfies an appropriate operationalisation of fairness. If algorithmic fairness is to promote trustworthiness and acceptability, such an operationalisation should be consistent with commonly shared intuitions and applicable societal and legal standards.

To facilitate the choice of appropriate fairness metrics, we develop a framework that makes explicit the normatively relevant assumptions behind different metrics.
For example, by requiring equal false negative and false positive rates for people of all ages, we implicitly assume that the risk associated with a false negative (or positive) is equally bad at all ages, or that it should not be considered. Neither of these assumptions is trivially true, but they are highly relevant to the ethical (and perhaps legal) assessment of the algorithm. However, for the medical diagnosis of a disease that affects older people more severely, one might instead justify unequal false-negative rates in favour of older age.
Since certain fairness metrics have been shown to be incompatible, we expect that in practice some metrics will regularly need to be prioritised over others, and this will also require justification based on normative rather than purely technical considerations.

Our framework is built upon the widely shared intuition, that *to be fair, we are required to treat like cases as like*, which has already been captured by Aristotle's formal equality principle. Fairness, therefore, does not generally mean treating everyone the same, but requires treating those cases differently that are different in a normatively relevant sense.
We begin our discussion with high-level normative questions including what 'like cases' actually are and what it means to treat cases 'as like'. Gradually, we delve into more specific questions, such as whether an algorithm's training data accurately reflects the 'ground truth' of case similarity, and which aspects are most critical when considering how benefits and risks should be distributed across groups, until we identify the technical specifications and statistical assumptions associated with each of the normative questions. In this way, we explore how certain answers constrain the choice of fairness metrics.
Conversely, given a fairness metric, our approach can be used to work out the normatively relevant assumptions that must hold to justify reliance on that metric.
Hence, our framework is a step towards structured and well-justified decision making about which fairness metric(s) to apply within a particular context.