Taming the AI Monster: Monitoring of Individual Fairness for Effective Human Oversight

Kevin Baum[1], Sebastian Biewer[2], Holger Hermanns[2], Sven Hetmank[3], Markus Langer[4], Anne Lauber-Rönsberg[3], and Sarah Sterz[2⋆]

[1] DFKI Saarbrücken, Neuro-Mechanistic Modeling
[2] Universität des Saarlandes, Department of Computer Science
[3] Technische Universität Dresden, IRGET
[4] Universität Freiburg, Department of Psychology

Abstract. This invited paper reviews a framework to assist in mitigating societal risks that software can pose. This is to promote effective human oversight, which is a central requirement enforced by the European Union's upcoming AI Act [29]. *The paper advertises fragments of an upcoming journal publication [12], and as such is itself low in genuine originality.* Yet it offers a specific perspective on that original work. Extrapolating earlier work on software doping, we report on the combination of established techniques for runtime monitoring and for probabilistic falsification to arrive at a black-box analysis technique for identifying undesired effects of software. We describe its application to high-risk systems that evaluate humans in a possibly unfair or discriminating way. The approach can assist humans-in-the-loop to make better informed and more responsible decisions. Our technical contribution is complemented by juridically, philosophically, and psychologically informed perspectives on the potential problems caused by AI systems.

Keywords: artificial intelligence, algorithmic fairness, probabilistic falsification, adequate trust, human oversight

1 Introduction

The lack of transparency of many AI-supported systems raises significant societal risks, including the potential for unfair or biased decision-making. This can lead to morally and instrumentally problematic outcomes, to breaches of legal obligations, to unfavourable societal effects, and to the undermining of public trust and acceptance of AI technologies. This is especially true for high-risk applications, which include credit approval [63], decisions on visa applications [54], admissions to higher education [85,19], screening of individuals in predictive policing [35], selection in HR [60,61,62], judicial decisions (as with COMPAS [21,23,3,47]), tenant screening [76], and more. In many of these areas, there are

⋆ Authors are listed alphabetically.

legitimate interests and valid reasons for using AI technology, although the risks associated with their use to date are manifold.

One frequently proposed remedy to the problems posed by the high-risk uses of opaque AI is *human oversight* [36,53,82] where a human expert is to make sure that the system operates in accordance with the desiderata set out by other human stakeholders. By now, the requirement for human oversight is even reflected in law, such as the *AI Act* of the European Union [29] that is about to be adopted or certain US state laws [84]. However, human oversight is not an unconditional remedy for any and all problems, and the effectiveness of a human overseer can be greatly reduced when certain conditions fail to be met. Notably, if the human overseer cannot gain enough knowledge about the system, their oversight will not achieve the desired aims. For example, if the overseer lacks any means to decide if a system made an unfair decision, they will have no reliable way of intervening when the system is, in fact, unfair. This is the problem that this paper has set out to tackle.

2 Setting the Stage

The challenge to overcome can best be introduced by an exemplary, albeit hypothetical admission system for higher education (inspired by [85,19]).

> *Example 1.* A large university assigns scores to applicants aiming to enter their computer science PhD program. The scores are computed using an automated, model-based procedure P, which is based on three data points: the position of the applicant's last graduate institution in an official, subject-specific ranking, the applicant's most recent grade point average (GPA), and their score in a subject-specific standardised test taken as part of the application procedure. The system then automatically computes a score for the candidate based on an estimation of how successful it expects them to be as students. A dedicated university employee, Unica is in charge of overseeing the individual outcomes of P and desk-reject candidates whose scores are below a certain, predefined threshold – unless she finds problems with P's scoring. The university pays especial attention to fairness in the scoring procedure, so Unica has to watch out for any signs of potential unfairness. If she suspects unfairness, Unica must decide on the case manually. Without any additional support, Unica, as human overseer in the loop, must manually check *all* cases for signs of unfairness as they are processed. This can be a tedious, complicated, and error-prone task and, as such, constitutes an impediment to the assumed scalability of the automated scoring process for high numbers of applicants. Therefore, she requires tools that assist her in detecting when something is off about the scoring of individual applicants.

Sometimes, we cannot mitigate all risks of high-risk AI in advance by technical measures, and some risk mitigation requires trade-off decisions involving features that are either impossible or difficult to operationalise and formalise. This is why

it is arguably essential that a human effectively oversees the system (which is also emphasised by several institutions such as UNESCO [82] and the *European High Level Expert Group* [36]), as well as in applicable law (such as the European AI Act [29] or the Washington State facial recognition law [84]). *Effective* human oversight, however, is only possible with the appropriate technical measures that allow human overseers to better understand the system at runtime [46,45]. From a technical point of view, this raises the pressing question of what such technical measures can and ought to look like to actually enable humans to live up to their responsibilities. Our contribution is intended to bridge the gap between the normative expectations of law and society and the current reality of technological design. Developing such a technical measure, a software tool supporting Unica, is thus the prime problem we focus on.

Our solution is based on the work developed in [12], in terms of a runtime monitor that provides automated assistance (based on [14]) to the human oversight and itself is based on a probabilistic falsification technique (introduced in [15]). All this is rooted in a suitable formal basis for rolling out runtime monitors for such high-risk systems that can detect and flag discrimination or unfair treatment of humans. We live up to the societal complexity of this example and provide an interdisciplinary situation analysis and an interdisciplinary assessment of the solution we shall propose.

The contributions echoed here from the original article [12] are twofold.

Promoting effective human oversight. We discuss and demonstrate a contribution to effective human oversight of high-risk systems, as required by the AI Act. The hypothetical university admission scenario introduced above will serve as a demonstrator for shedding light on the applicability of our approach and on the principles behind it. On a conceptual level, we consider it important to clarify which duties come with the usage of such a system; from a *legal* perspective, particularly considering the AI Act, substantiated by considering the *ethical* dimension from a philosophical perspective, and from a *psychological* perspective, particularly deliberating on how the overseeing can become *effective*.

Falsification-based test input generation. On a technical level, we describe how recent work [13] on a formal framework for robust cleanness can be combined with a probabilistic falsification technique to identify problems of fairness and discrimination in AI usages akin to the admission scenario described above. We describe a search procedure that aims at generating synthetic data of (hypothetical) applicants whose parameters are very similar to the individuals currently looked at but who are classified differently by the AI. The approach uses a fairness test procedure, and the problem then is to effectuate test input selection in a meaningful manner. In this, probabilistic falsification supports the testing procedure by guiding it towards test inputs that make the fairness tests likely to fail. Altogether, we arrive at a runtime monitor for individual fairness based on probabilistic falsification. This we consider as a core component for assisting humans who need to oversee scenarios as the one described above.

While the contents of this paper are subsumed by the original contents of [12], the former has been rearranged in order to directly put in focus the use of the above contributions for the benefit of human oversight. In this respect, this paper offers a distinct value to the interested reader.

3 Fairness, Discrimination, Explainability

Our contribution draws on and adds to three vibrant topics of current research, namely *Explainable AI (XAI)*, *AI Fairness*, and *Discrimination*.

Explainable AI. Many of the most successful AI systems today are black boxes of some kind [8]. Accordingly, the field of "Explainable AI" [32] focuses on the question of how to provide users (and possibly other stakeholders) with more information via several key perspicuity properties [78] of these systems and their outputs to make them understand these systems and their outputs in ways necessary to meet various desiderata [55,49,44,4,59,20]. The concrete expectations and promises associated with various XAI methods are manifold. Among them are enabling warranted trust in systems [69,73,39,42,9], increasing human-system decision-making performance [43], for instance through increasing human situation awareness when operating systems [71], enabling responsible decision-making and effective human oversight [10,51,75], as well as identifying and reducing discrimination [49]. It often remains unclear what kind of explanations are generated by the various explainability methods and how they are meant to contribute to the fulfilment of the desiderata, even though these questions have become the subject of systematic and interdisciplinary research [46,44].

Our approach can be taxonomised along at least two different distinctions [69,68,56,46,77]: First, it is *model-agnostic* (not *model-specific*), i.e., it is not tailored to a particular class of models but operates on observable behaviour – the inputs and outputs of the model. Second, our method is a *local method* (not *global*), i.e., it is meant to shed light on certain outputs rather than the system as a whole.

Fairness. Fairness, discrimination, justice, equal opportunity, bias, prejudice, and many more such concepts are part of a meaningfully interrelated cluster that has been analysed and dissected for millennia [5,6]. Many fields are traditionally concerned with the concepts of fairness and discrimination, ranging from philosophy [5,6,25,31,66,65,67] to legal sciences [18,83,34,81], to psychology [37,88], to sociology [2,40], to political theory [66], to economics [33]. Nowadays, it has also become a technological topic that calls for cross-disciplinary perspectives [30]. It is widely recognised that discrimination by unfair classification and regression models is one particularly important risk of AI-supported decision making. As a result, a colourful zoo of different operationalisations of unfairness has emerged [83,64], which should be seen less as a set of competing approaches and more as mutually complementary [31].

With regard to fairness, two distinctions are especially relevant to our work. First, one distinction is made between *individual fairness*, i.e., that similar individuals are treated similarly [24], and *group fairness*, i.e., that there is adequate group parity [16]. Measures of individual fairness are often close to the Aristotelian dictum to treat like cases alike [5,6]. In a sense, operationalisations of individual fairness are robustness measures [79,17], but instead of requiring robustness with respect to noise or adversarial attacks, measures of individual fairness, such as the one by Dwork et al. [24], call for robustness with respect to highly context-dependent differences between representations of human individuals. Second, recent work from the field of law [83] suggests to differentiate between *bias preserving* and *bias transforming* fairness metrics. Bias preserving fairness metrics seek to avoid adding new bias. For such metrics, historic performances are the benchmarks for models, with equivalent error rates for each group being a constraint. In contrast, bias transforming metrics do not accept existing bias as a given or neutral starting point but aim at adjustment. Therefore, they require to make a "positive normative choice" [83], i.e., to actively decide which biases the system is allowed to exhibit and which it must not exhibit.

Over the years, many concrete approaches have been suggested to foster different kinds of fairness in artificial systems, especially in AI-based ones [52,49,83,86,64]. Yet, to the best of our knowledge, an approach like ours is still missing. One of the approaches that are closest to ours, namely that by John et al. [41], is not local and, therefore, not suitable for runtime monitoring. Also, it is not model-agnostic. So, to the best of our knowledge, our approach provides a new contribution to the debate on unfairness detection.

It is important to note/recognise that our approach can only be understood as part of a more holistic approach to preventing or reducing unfairness. After all, there are many sources of unfairness [7] (also see Figure 1). Therefore, not every technical measure can detect every kind of unfairness, and eliminating one source of unfairness might not be sufficient to eliminate all unfairness. Our approach tackles only unfairness introduced by the system, but not other kinds of unfairness.

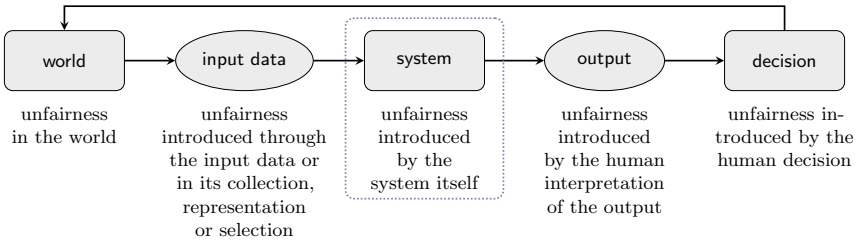


Fig. 1: Sketch of different origins of unfairness in a decision process supported by a system; the dotted box indicates which unfairness our monitoring targets.

Discrimination. We understand discrimination as dissimilar treatment of similar cases or similar treatment of dissimilar cases without justifying reason. This definition can also be found in the law [28, §43]. Our work exclusively focuses on discrimination *qua* dissimilar treatment of similar cases. Discrimination requires a thoughtful and largely not formalisable consideration of "justifying reason." However, we will exploit the relation between discrimination and fairness: Unfairness in a system can arguably be a good proxy of discrimination – even though not every unfair treatment by a system necessarily constitutes discrimination (especially not in the legal sense). Thus, a tool that highlights cases of unfairness in a system can be highly instrumental in detecting discriminatory features of a system. It is not viable, though, to let such a tool rule out unfair treatment fully automatically without human oversight since there could be justifying reasons to treat two similar inputs in a dissimilar way.

4 Individual Fairness of Systems Evaluating Humans

Against the contextual backdrop as described above, we now return to the characteristics of Example 1, where Unica uses an AI system that is supposed to assist her with the selection of applicants for a hypothetical university. A usable fairness analysis can happen no later than at runtime since Unica needs to make a timely decision on whether to include the applicant in further considerations. We describe technical measures that help mitigate this challenge by providing her with information from an individual fairness analysis in a suitable, purposeful, and expedient way. To this end, we propose a formal definition for individual fairness extending the one by Dwork et al. [24], extrapolating earlier work on robust cleanness [11]. We develop a runtime monitor that analyses every output of P immediately after P's decision, which strategically searches for unfair treatment of a particular individual by comparing them to relevant hypothetical alternative individuals so as to provide a fairness assessment in a timely manner.

4.1 Individual Fairness

Unica from Example 1 should be able to detect individual unfairness. An operationalisation thereof by Dwork et al. [24] is based on the Lipschitz condition to enforce that similar individuals are treated similarly. To measure similarity, they assume the existence of distance functions measuring the distances between system inputs as well as between system outputs. A function $d : X \times X \to \overline{\mathbb{R}}_{\geq 0}$ is a *distance function* if and only if it satisfies $d(x, x) = 0$ and $d(x, y) = d(y, x)$. In this, $\overline{\mathbb{R}}_{\geq 0} \coloneqq \{x \in \mathbb{R} \mid x \geq 0\} \cup \{\infty\}$ is the set of the non-negative extended real numbers. Two such functions are assumed to exist, namely d_{In} operating on the set In of system inputs, and similarly d_{Out} for system output (from set Out). In the example, the inputs are the data vectors representing human individuals, and outputs correspond to the scores produced.

Dwork et al. [24] assume that both distance functions perfectly measure distances between individuals[5] and between outputs of the system, respectively, but admit that in practice these distance functions are only approximations of a ground truth at best. They suggest that distance measures might be learned, but there is no one-size-fits-all approach to selecting distance measures. Indeed, obtaining such distance metrics is a topic of active research [87,57,38]. Additionally, the Lipschitz condition assumes a Lipschitz constant L to establish a linear constraint between input and output distances.

Definition 1. *A program* $\mathsf{P} : \mathsf{In} \to \mathsf{Out}$ *is* Lipschitz-fair *w.r.t.* $d_{\mathsf{In}} : \mathsf{In} \times \mathsf{In} \to \mathbb{R}$, $d_{\mathsf{Out}} : \mathsf{Out} \times \mathsf{Out} \to \mathbb{R}$, *and a Lipschitz constant* L, *if and only if for all* $\mathsf{i}_1, \mathsf{i}_2 \in \mathsf{In}$, $d_{\mathsf{Out}}(\mathsf{P}(\mathsf{i}_1), P(\mathsf{i}_2)) \leq L \cdot d_{\mathsf{In}}(\mathsf{i}_1, \mathsf{i}_2)$.

Lipschitz-fairness comes with some restrictions that limit its suitability for practical application:

d_{In}-d_{Out}-relation: High-risk systems are typically complex systems and ask for more complex fairness constraints than the linearly bounded output distances provided by the Lipschitz condition. For example, using the Lipschitz condition prevents us from allowing small local jumps in the output and, at the same time, from forbidding jumps of the same rate of increase over larger ranges of the input space.

Input relevance: The condition quantifies over the entire input domain of a program. This overlooks two things: first, it is questionable whether each input in such a domain is plausible as a representation of a real-world individual. But whether a system is unfair for two implausible and purely hypothetical inputs is largely irrelevant in practice. Secondly, it also ignores that mere potential unfair treatment is at most a threat, not necessarily already a harm [70]. Therefore, even with a restriction to only plausible applicants, the analysis might take into account more inputs than needed for many real-world applications. What is important in practice is the ability to determine whether *actual* applicants are treated unfairly – and for this, it is often not needed to look at the entire input domain.

Monitorability: In a monitoring scenario with the Lipschitz condition in place, a fixed input i_1 must be compared to potentially all other inputs i_2. Since the input domain of the system can be arbitrarily large, the Lipschitz condition is not yet suitable for monitoring in practice (for a related point see John et al. [41]).

We propose a notion of individual fairness that, instead of the constant L, uses a function f to relate input distances and output distances in a more general way.

[5] For easier readability, we will not distinguish between *individuals* and their *representations* unless this distinction is relevant in the specific context. It is nevertheless important to note that inputs are not individuals, but only representations of individuals, since an input could inadequately represent an individual and therefore be unfair.

Further, we make it explicit that d_{In}, d_{Out}, and f are parameters of the fairness notion by encapsulating them in triples $\mathcal{F} = \langle d_{\mathsf{In}}, d_{\mathsf{Out}}, f \rangle$ that we call *fairness contracts*. Our fairness definition evaluates fairness for a finite set of individuals $\mathcal{I} \subseteq \mathsf{In}$ (e.g., a set of applicants). A fairness contract specifies certain fairness parameters for a concrete context or situation. Such parameters should generally not already include $\mathcal{I}$ to avoid introducing new unfairness through the monitor by tailoring it to specific inputs individually or by treating certain inputs differently from others. We can operationalise[6] individual fairness as follows:

Definition 2. *A program* $\mathsf{P} : \mathsf{In} \to \mathsf{Out}$ *is* individually fair *for a set* $\mathcal{I} \subseteq \mathsf{In}$ *of actual inputs w.r.t. a fairness contract* $\mathcal{F} = \langle d_{\mathsf{In}}, d_{\mathsf{Out}}, f \rangle$*, if and only if for all* $\mathsf{i} \in \mathcal{I}$ *and all* $\mathsf{i}' \in \mathsf{In}$*,* $d_{\mathsf{Out}}(\mathsf{P}(\mathsf{i}), \mathsf{P}(\mathsf{i}')) \leq f(d_{\mathsf{In}}(\mathsf{i}, \mathsf{i}'))$*.*

The idea behind *individual fairness* is that every individual in set $\mathcal{I}$ is compared to potential other inputs in the domain of P. These other inputs do not necessarily need to be in $\mathcal{I}$, nor do these inputs need to have "physical counterparts" in the real world. Driven by the insights of the *Input relevance* restriction of Lipschitz-fairness, we explicitly distinguish inputs in the following and will call inputs that are given to P by a user *actual inputs*, denoted i_a, and call inputs to which such i_a are compared *synthetic inputs*, denoted i_s. Actual inputs typically[7] are inputs that have a real-world counterpart, while this might or might not be true for synthetic inputs.

At first glance, it might seem sufficient to use only actual inputs. This way, for example, Unica would find out whether one applicant was treated unfairly relative to another applicant. This, however, is not enough: Unica is after *any* unfairness of the system towards a certain applicant, and not just one relative to some other, actual applicant. At the same time, Unica cannot and should not expect that, coincidentally, a candidate has applied who has the very specific properties needed to unveil the system's unfairness towards another candidate. Hence, synthetic inputs are worthwhile to be considered.

Notice that *individual fairness* is a conservative extension of Lipschitz-fairness. With $\mathcal{I} = \mathsf{In}$ and $f(x) = L \cdot x$, *individual fairness* mimics Lipschitz-fairness. Wachter et al. [83] classify the Lipschitz-fairness of Dwork et al. [24] as bias-transforming. As we generalise this and introduce no element that has to be regarded as bias-preserving, our approach arguably is bias-transforming, too.

Individual fairness, with its function f, provides a powerful tool to model complex fairness constraints. How such an f is defined has a profound impact

[6] Definition 2 is not a *definition* of individual fairness in the strict sense, since individual fairness already has a meaning, namely that similar individuals are treated similarly, as described above in Section 3. It rather is an *operationalisation* that has to be employed appropriately in order to yield a proper measure of individual fairness. This, for example, includes a parameterisation with a fitting fairness contract that is meaningful in the context of individual fairness, and fixes what similarity is to mean in this context. Nevertheless, Definition 2 is a suitable operationalisation for our purposes. In this paper, it will be clear from the context whether we will talk about individual fairness in its original sense or in terms of the operationalisation.

[7] A case where actual inputs might not have real-world counterparts is testing.

Algorithm 1 FairnessMonitor,
with ξ-min $S = (\xi, \mathsf{i}_1, \mathsf{i}_2)$ only if $(\xi, \mathsf{i}_1, \mathsf{i}_2) \in S$ and for all $(\xi', \mathsf{i}'_1, \mathsf{i}'_2) \in S$, $\xi' \geq \xi$

Falsification Parameters: PS: Proposal scheme, β: Temperature parameter
Input: System $P : \mathsf{In} \to \mathsf{Out}$, Fairness contract $\mathcal{F} = \langle d_{\mathsf{In}}, d_{\mathsf{Out}}, f \rangle$, and set of actual inputs $\mathcal{I}$
Output: A minimal fairness score triple from $\mathbb{R} \times \mathcal{I} \times \mathsf{In}$.

$\mathsf{i}_\mathsf{s} \leftarrow$ any input $\mathsf{i}_\mathsf{a} \in \mathcal{I}$
$(\xi, \mathsf{i}_{\mathsf{min}}, \mathsf{i}_\mathsf{s}) \leftarrow \xi\text{-}\mathsf{min}\{(F(\mathsf{i}_\mathsf{a}, \mathsf{i}_\mathsf{s}), \mathsf{i}_\mathsf{a}, \mathsf{i}_\mathsf{s}) \mid \mathsf{i}_\mathsf{a} \in \mathcal{I}\}$
$(\xi_{\mathsf{min}}, \mathsf{i}_1, \mathsf{i}_2) \leftarrow (\xi, \mathsf{i}_{\mathsf{min}}, \mathsf{i}_\mathsf{s})$
while not timeout **do**
  $\mathsf{i}'_\mathsf{s} \leftarrow \mathsf{PS}(\mathsf{i}_\mathsf{s}, \mathsf{P}(\mathsf{i}_\mathsf{s}))$
  $(\xi', \mathsf{i}'_{\mathsf{min}}, \mathsf{i}'_\mathsf{s}) \leftarrow \xi\text{-}\mathsf{min}\{(F(\mathsf{i}_\mathsf{a}, \mathsf{i}'_\mathsf{s}), \mathsf{i}_\mathsf{a}, \mathsf{i}'_\mathsf{s}) \mid \mathsf{i}_\mathsf{a} \in \mathcal{I}\}$
  $(\xi_{\mathsf{min}}, \mathsf{i}_1, \mathsf{i}_2) \leftarrow \xi\text{-}\mathsf{min}\{(\xi_{\mathsf{min}}, \mathsf{i}_1, \mathsf{i}_2), (\xi', \mathsf{i}'_{\mathsf{min}}, \mathsf{i}'_\mathsf{s})\}$
  $\alpha \leftarrow \exp(-\beta(\xi' - \xi))$
  $r \leftarrow \mathsf{UniformRandomReal}(0, 1)$
  if $r \leq \alpha$ **then**
    $\mathsf{i}_\mathsf{s} \leftarrow \mathsf{i}'_\mathsf{s}$
    $\xi \leftarrow \xi'$
  end if
end while
return $(\xi_{\mathsf{min}}, \mathsf{i}_1, \mathsf{i}_2)$

on the quality of the fairness analysis. A full discussion about which types of functions make a good f goes beyond the scope of this paper. What are suitable choices for f and the distance functions d_{In} and d_{Out} heavily depends on the context in which fairness is analysed – there is no one-fits-it-all solution. *Individual fairness* makes this explicit with the formal fairness contract $\mathcal{F} = \langle d_{\mathsf{In}}, d_{\mathsf{Out}}, f \rangle$.

4.2 Fairness Monitoring

We now develop a fairness monitor based on probabilistic falsification [1]. Given a set of actual inputs, the monitor searches for a synthetic counterexample to falsify a system P w.r.t. a fairness contract $\mathcal{F}$. To this end, we define a *fairness score* as function $F(\mathsf{i}_\mathsf{a}, \mathsf{i}_\mathsf{s}) \coloneqq f(d_{\mathsf{In}}(\mathsf{i}_\mathsf{a}, \mathsf{i}_\mathsf{s})) - d_{\mathsf{Out}}(\mathsf{P}(\mathsf{i}_\mathsf{a}), \mathsf{P}(\mathsf{i}_\mathsf{s}))$. With regard to probabilistic falsification F is a quantitative description of *individual fairness* that serves as a robustness estimate. That is, if $F(\mathsf{i}_\mathsf{a}, \mathsf{i}_\mathsf{s})$ is non-negative, then $d_{\mathsf{Out}}(\mathsf{P}(\mathsf{i}_\mathsf{a}), \mathsf{P}(\mathsf{i}_\mathsf{s})) \leq f(d_{\mathsf{In}}(\mathsf{i}_\mathsf{a}, \mathsf{i}_\mathsf{s}))$, and if it is negative, then $d_{\mathsf{Out}}(\mathsf{P}(\mathsf{i}_\mathsf{a}), \mathsf{P}(\mathsf{i}_\mathsf{s})) \not\leq f(d_{\mathsf{In}}(\mathsf{i}_\mathsf{a}, \mathsf{i}_\mathsf{s}))$. For a set of actual inputs $\mathcal{I}$, the definition generalises to $F(\mathcal{I}, \mathsf{i}_\mathsf{s}) \coloneqq \min\{F(\mathsf{i}_\mathsf{a}, \mathsf{i}_\mathsf{s}) \mid \mathsf{i}_\mathsf{a} \in \mathcal{I}\}$, i.e., the overall fairness score is the minimum of the concrete fairness scores of the inputs in $\mathcal{I}$.

Algorithm 1 shows FairnessMonitor, which searches for the minimal fairness score in a system P for fairness contract $\mathcal{F}$. The algorithm stores fairness scores in triples that also contain the two inputs for which the fairness score was computed. The minimum in a set of such triples is defined by the function ξ-min that returns the triple with the smallest fairness score of all triples in the set. The first line of FairnessMonitor initialises the variable i_s with an arbitrary actual input from $\mathcal{I}$.

For this value of $\mathsf{i_s}$, the algorithm checks the corresponding fairness scores for all actual inputs $\mathsf{i_a} \in \mathcal{I}$ and stores the smallest one. In line 3, the globally smallest fairness score triple is initialised. In line 5, the algorithm uses a parameterisable proposal scheme to get the next synthetic input $\mathsf{i'_s}$. Line 6 is similar to line 2: for the newly proposed $\mathsf{i'_s}$ it finds the smallest fairness score, stores it, and updates the global minimum if it found a smaller fairness score (line 7). Lines 8-13 are the heart of the probabilistic search for an example that violates fairness; it comes from the original algorithm proposed by Abbas et al. [1]. Our variant of the algorithm does not (exclusively) aim to falsify the fairness property, but aims at minimising the fairness score; even if the fair treatment of the inputs in $\mathcal{I}$ cannot be falsified in a reasonable amount of time, we still learn how robustly they are treated fairly, i.e., how far the least fairly treated individual in $\mathcal{I}$ is away from being treated unfairly. After the timeout occurs, the algorithm returns the triple with the overall smallest seen fairness score ξ_{min}, together with the actual input i_1 and the synthetic input i_2 for which ξ_{min} was found. In case ξ_{min} is negative, i_2 is a counterexample for P being individually fair.

FairnessMonitor implements a sound $\mathcal{F}$-unfairness detection as stated in Proposition 1. However, it is not complete, i.e., it is not generally the case that P is individually fair for $\mathcal{I}$ if ξ is positive. It may happen that there is a counterexample, but FairnessMonitor did not succeed in finding it before the timeout.

Proposition 1. *Let* $\mathsf{P} : \mathsf{In} \to \mathsf{Out}$ *be a program,* $\mathcal{F} = \langle d_{\mathsf{In}}, d_{\mathsf{Out}}, f \rangle$ *a fairness contract, and* $\mathcal{I}$ *a set of actual inputs. Further, let* $(\xi_{\mathsf{min}}, \mathsf{i}_1, \mathsf{i}_2)$ *be the result of* $\mathsf{FairnessMonitor}(\mathsf{P}, \mathcal{F}, \mathcal{I})$*. If* ξ_{min} *is negative, then* P *is not individually fair for* $\mathcal{I}$ *w.r.t.* $\mathcal{F}$*.*

Moreover, FairnessMonitor circumvents major restrictions of Lipschitz-fairness:

d_{In}-d_{Out}-relation: *Individual fairness* defines constraints between input and output distances by means of a function f, which allows the expression of complex fairness constraints. For a more elaborate discussion, see [12, Appendix A].

Input relevance: *Individual fairness* explicitly distinguishes between actual and synthetic inputs. This way, *individual fairness* acknowledges a possible obstacle of the fairness theory when it comes to real-world usage of the analysis, namely that only some elements of the system's input domain might be plausible, and usually, only a few of them become actual inputs that have to be monitored for unfairness.

Monitorability: FairnessMonitor demonstrates that *individual fairness* is monitorable. It resolves the quantification over In using the above concepts from probabilistic falsification using the robustness estimate function F as defined above.

Towards individual fairness in the loop. If a high-risk system is in operation, a human in the loop must oversee the correct and fair functioning of the outputs of the system. To do this, the human needs real-time fairness information. Figure 2 shows how this can be achieved by coupling the system P and the FairnessMonitor

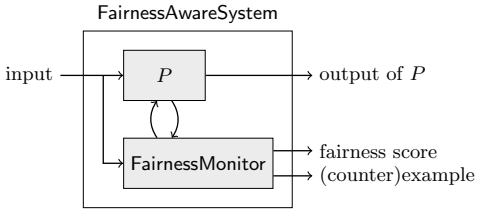


Fig. 2: Schematic visualisation of FairnessAwareSystem

in Algorithm 1 in a new system called FairnessAwareSystem. FairnessAwareSystem is sketched in Algorithm 2.

Algorithm 2 FairnessAwareSystem

Parameters: System $\mathsf{P} : \mathsf{In} \to \mathsf{Out}$, Fairness contract $\mathcal{F} = \langle d_{\mathsf{In}}, d_{\mathsf{Out}}, f \rangle$
Input: Input $\mathsf{i_a} \in \mathsf{In}$
Output: Tuple of the system output, normalised fairness score, and synthetic values witnessing the fairness score

1: $(\xi_{\mathsf{min}}, \mathsf{i_a}, \mathsf{i_s}) \leftarrow \mathsf{FairnessMonitor}(\mathsf{P}, \mathcal{F}, \{\mathsf{i_a}\})$
2: **return** $\left(\mathsf{P}(\mathsf{i_a}), \frac{\xi_{\mathsf{min}}}{f(d_{\mathsf{In}}(\mathsf{i_a}, \mathsf{i_s}))}, (\mathsf{i_s}, \mathsf{P}(\mathsf{i_s}))\right)$

Intuitively, the FairnessAwareSystem is a higher-order program that is parameterised with the original program P and the fairness contract $\mathcal{F}$. When instantiated with these parameters, the program takes arbitrary (actual) inputs $\mathsf{i_a}$ from In. In the first step, it does a fairness analysis using FairnessMonitor with arguments P, $\mathcal{F}$, and $\{\mathsf{i_a}\}$. To make fairness scores comparable, FairnessAwareSystem normalises the fairness score ξ received from FairnessMonitor by dividing[8] it by the output distance limit $f(d_{\mathsf{In}}(\mathsf{i_a}, \mathsf{i_s}))$. For fair outputs, the score will be between 0 (almost unfair) and 1 (as fair as possible).[9] Outputs that are not individually fair are accompanied by a negative score representing how much the limit $f(d_{\mathsf{In}}(\mathsf{i_a}, \mathsf{i_s}))$ is exceeded. A fairness score of $-n$ means that the output distance of $\mathsf{P}(\mathsf{i_a})$ and $\mathsf{P}(\mathsf{i_s})$ is $n + 1$ times as high as that limit. Finally, FairnessAwareSystem returns the triple with P's output for $\mathsf{i_a}$, the normalised fairness score, and the synthetic input with its output witnessing the fairness score.

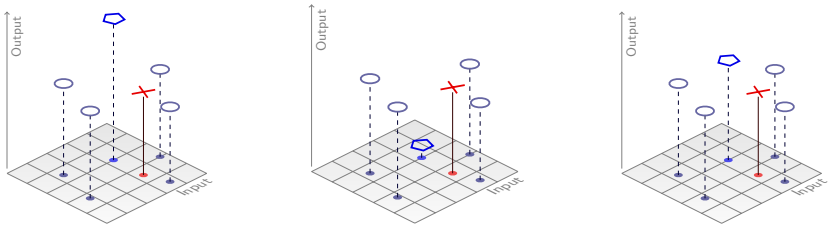
[8] For f that can return 0, a division of zero by zero may occurr. The result of this division should be defined depending on the concrete context; reasonable values range from the extreme scores 0 (to indicate that the score is on edge of becoming 'unfair') to 1 (to indicate that more fairness is impossible).

[9] Fairness may be a vague concept that cannot be dichotomised. By its choice of the fairness contract parameters, our approach nevertheless specifies a (non-arbitrary) cut-off point at 0; but it does so for purely instrumental and non-ontological reasons.

Interpretation of monitoring results. Especially when FairnessAwareSystem finds a violation of *individual fairness*, the suitable interpretation and appropriate response to the normalised fairness score proves to be a non-trivial matter that requires expertise.

> *Example 2.* Instead of using P from Example 1 on its own, Unica now uses FairnessAwareSystem with a suitable fairness contract. and thereby receive a fairness score along with P's verdict on each applicant. (Which fairness contracts are suitable is an open research problem, see *Limitations & Challenges* in Section 6.) If the fairness score is negative, she can also take into account the information on the synthetic counterpart returned by FairnessAwareSystem. Among the 4096 applicants for the PhD program, the monitoring assigns a negative fairness score to three candidates: Alexa, who received a low score, Eugene, who was scored very highly, and John, who got an average score. According to their scoring, Alexa would be desk-rejected, while Eugene and John would be considered further.
> Alexa's synthetic counterpart, let's call him Syntbad, is ranked much higher than Alexa. In fact, he is ranked so high that Syntbad would not be desk-rejected. Unica compares Alexa and Syntbad and finds that they only differ in one respect: Syntbad's graduate university is the one in the official ranking that is immediately *below* the one that Alexa attended. Unica does some research and finds that Alexa's institution is predominantly attended by People of Colour, while this is not the case for Syntbad's institution. Therefore, FairnessAwareSystem helped Unica not only to find an unfair treatment of Alexa, but also to uncover a case of potential racial discrimination.
> John's counterpart, Synclair, is ranked much lower than him. Unica manually inspects John's previous institution (an infamous online university), his GPA of 1.8, and his test result with only 13%. She finds that this very much suggests that John will not be a successful PhD candidate and desk-rejects him. Therefore, Unica has successfully used FairnessAwareSystem to detect a fault in the scoring system P whereby John would have been treated unfairly in a way that would have been to his advantage.
> Eugene received a top score, but his synthetic counterpart, Syna, received only an average one. Unica suspects that Eugene was ranked too highly given his graduate institution, GPA, and test score. However, as he would not have been desk-rejected either way, nothing changes for Eugene, and the unfairness he was subject to, is not of effect to him.
> The cases of John and Eugene share similarities with the configuration in (b) in Figure 3, the one of Alexa with (a), and the ones of all other 4093 candidates with (c).

If our monitor finds only a few problematic cases in a (sufficiently large and diverse) set of inputs, our monitoring helps Unica from our running example by drawing her attention to cases that require special attention. Thereby, individuals who are judged by the system have a better chance of being treated fairly, since



(a) case of unfairness where input is treated worse than relevant counterpart

(b) case of unfairness where input is treated better than relevant counterpart

(c) case of no detected unfairness

Fig. 3: Exemplary illustration of configurations of an input (red cross) and its synthetic counterparts (grey circles) and the synthetic counterpart with the minimal fairness score (blue polygon); with a two-dimensional input space (grid) and a one-dimensional output.

even rare instances of unfair treatment can be detected. If, on the other hand, the number of problematic cases found is large or Unica finds especially concerning cases or patterns, this can point to larger issues within the system. In these cases, Unica should take appropriate steps and make sure that the system is no longer used until clarity is established as to why so many violations or concerning patterns are found. If the system is found to be systematically unfair, it should arguably be removed from the decision process. A possible conclusion could also be that the system is unsuitable for certain use cases, e.g., for the use of individuals from a particular group. Accordingly, it might not have to be removed altogether but only needs to be restricted such that problematic use cases are avoided. In any case, significant findings should also be reported to the developers or deployers of the potentially problematic system. A fairness monitoring such as in FairnessAwareSystem or a fairness analysis as in FairnessMonitor could also be useful to developers, regulating authorities, watchdog organisations, or forensic analysts as it helps them to check the individual fairness of a system in a controlled environment.

Remark 1. Individual fairness is called func-fairness in [12] and is an adaptation of func-cleanness, which has been studied in earlier work [22,11] on *software doping*. In this context, a cleanness property – like func-cleanness – characterises the absence of doped software. Intuitively, software doping relates to the existence of a hidden feature in a software that was added intentionally by the software manufacturer, but which is not in the interest of the user or society. The diesel emissions scandal is by now the archetypal example of software doping: various car manufacturers added defeat devices into their emission cleaning systems to distinguish whether the car is undergoing an emissions test from whether it is used in normal operation on the road. In the former case, the emission cleaning worked as required, while in the latter case, the engine control system was optimising

for other objectives instead, thereby effectively infringing legal requirements. A falsification-based monitoring approach, including a logical characterisation of various notions of cleanness, has been developed for the diesel use-case [15,12].

5 Interdisciplinary Assessment of Fairness Monitoring

The upcoming AI act stresses the need for human oversight of AI systems, but its stipulations are not free of ambiguities and the need for interpretation. This raises the question of whether our approach meets requirements that go beyond pre-theoretical deliberations. We here assess some key normative aspects in philosophical and legal terms, and also briefly turn to the related empirical aspects, especially from psychology.

5.1 Psychological assessment

Fairness monitoring promises various advantages in terms of human-system interaction in application contexts – provided it is extended by an adequate user interface – which calls for empirical tests and studies. We will only discuss a possible benefit that closely aligns with the upcoming AI Act: our approach may support effective human oversight. Two central aspects of effective oversight are situation awareness and warranted trust. Our method highlights unfairness in outputs which can be expected to increase users' situation awareness (i.e., "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" [26, p. 36]), which is a variable central for effective oversight [27]. In the minimal case, this allows users to realise that something requires their attention and that they should check the outputs for plausibility and adequacy. In the optimal case and after some experience with the monitor, it may even allow users to predict instances where a system will produce potentially unfair outputs. In any case, the monitoring should enable them to understand the limitations of the system and to feed back their findings to developers who can improve the system. This leads us to warranted trust, which includes that users are able to adequately judge when to rely on system outputs and when to reject them [48,39]. Building warranted trust strongly depends on users being able to assess system trustworthiness in the given context of use [72,48]. According to their theoretical model on trust in automation, Lee and See [48] propose that trustworthiness relates to different facets of which performance (e.g., whether the system performs reliably with high accuracy) and process (e.g., knowing how the system operates and whether the system's decision-processes help to fulfil the trustor's goals) are especially relevant in our case. Specifically, fairness monitoring should enable users to judge system performance more accurately (e.g., by revealing possible issues with system outputs) and system processes (e.g., whether the system's decision logic was appropriate). In line with Lee and See's propositions, this should provide a foundation for users to judge system trustworthiness better and should thus be a promising means to promote warranted trust. In consequence,

our monitoring provides a needed addition to high-risk use contexts of AI because it offers information enabling humans to more adequately use AI-based systems in the sense of possibly better human-system decision performance and with respect to user duties as described in the AI Act.

5.2 Philosophical assessment

More effective oversight promises more informed decision-making. This, in turn, enables morally better decisions and outcomes since humans can morally ameliorate outcomes in terms of fairness and can see to it that moral values are promoted. Fairness monitoring also helps safeguard fundamental democratic values if it is applied to potentially unfair systems used in certain societal institutions of a high-risk character, such as courts or parliaments. It could, for example, make AI-aided court decisions more transparent and promote equality before the law. However, since our approach requires finding context-appropriate and morally permissible parameters for $\mathcal{F}$, moral requirements arise to enable the finding of such parameters. This affects not only developers of such systems but also those who are in a position to enforce that adequate parameters are chosen, such as governmental authorities, supervising institutions, or certifiers.

Apart from that, various parties have arguably a legitimate interest in adequately ascribing moral responsibility for the outcomes of certain decisions to human deciders [10] – regardless of whether the decision-making process is supported by a system. Adequately ascribing moral responsibility is not always possible, though. One precondition for moral responsibility is that the agent had sufficient epistemic access to the consequences of their doing [80,58], i.e., that they have enough and sufficiently well-justified beliefs about the results of their decision. Someone overseeing a university selection process (like Unica) should, for example, have sufficiently well-justified beliefs that, at the very least, their decisions do not result in more unfairness in the world. If the admission process is supported by a black-box system, though, Unica cannot be expected to have any such beliefs since she lacks insight in the fairness of the system. Therefore, adequate responsibility ascription is usually not possible in this scenario. Our monitoring alleviates this problem by providing the decider with better epistemic access to the fairness of the system.

FairnessAwareSystem helps in making Unica's role in the decision process significant and not only that of a mere button-pusher. FairnessAwareSystem makes it possible for her to fulfil some of the responsibilities and duties plausibly associated with her role. For example, she can now be realistically expected to not only detect, but resolve at least some cases of apparent unfairness competently (although she may need additional information to do so). In this respect, she should not be 'automated away' (cf. [50]).

5.3 Legal assessment

A central legislative debate of our time is how to counter the risks AI systems can pose to the health and safety or fundamental rights of natural persons.

Protective measures must be taken at various levels: First, before being permitted on the market, it must be ensured *ex-ante* that such high-risk AI-systems are in conformity with mandatory requirements regarding safety and human rights [29, Art. 16, Art 27]. This means in particular that the selection of the properties that a system should exhibit requires a positive normative choice and should not simply replicate biases present in the status quo [83]. In addition, AI-systems must be designed and developed in such a way that natural persons can oversee their functioning. For this purpose, it is necessary for the provider to design and develop the AI system in such a way that it includes appropriate features enabling human oversight before it is placed on the market or put into service [29, Art. 14].

Second, during runtime, the proper functioning of high-risk AI systems that have been legally placed on the market must be ensured. To achieve this goal, a bundle of different measures is needed, ranging from legal obligations to implement and perform meaningful oversight mechanisms to user training and awareness in order to counteract 'automation bias'. In particular, such measures should guarantee that the natural persons to whom human oversight has been assigned have the necessary competence, training, and authority to carry out that role [29, Art. 26 (2)]. Furthermore, the AI Act proposal requires deployers to inform the provider or distributor and suspend the use of the system when they have identified any serious incidents or any malfunctioning [29, Art. 26(5)].

Third, and *ex-post*, providers must act and take the necessary corrective actions as soon as they become aware, e.g., through information provided by the deployer, that the high-risk system does not (or no longer) meet the legal requirements [29, Art. 20]. To this end, they must establish and document a system of monitoring that is proportionate to the type of AI technology and the risks of the high-risk AI system [29, Art. 72].

Fairness monitoring can be helpful in all three of the above respects. Therefore, we argue that there is even a legal obligation to use technical measures such as the method presented in this paper if this is the only way to ensure effective human oversight.

6 Conclusion

This invited paper has echoed elements of a forthcoming journal publication [12] that applies runtime monitoring and probabilistic falsification techniques to high-risk (AI) systems.

We have looked at a runtime fairness monitor to promote effective human oversight of high-risk systems. An interdisciplinary evaluation from a psychological, philosophical, and legal perspective complements the development of this monitor. As seen in Figure 1, our fairness monitoring aims to uncover a particular kind of unfairness, namely individual unfairness, that originates from within the system. This does not include group unfairness as well as unfairness from sources other than the system. Another limitation is the need to account for the human's competence to interpret the system outputs. Even though this is

not a limitation that is inherent to our approach, it nevertheless will arguably be relevant in some practical cases, and an implementation of the monitoring always has to happen with the human in mind. For example, the design of the tool should avoid creating the false impression that the system is proven to be fair for an individual if no counterexample has been found. Interpretations like this could lead to inflated judgements of system trustworthiness and eventually to overtrusting system outputs [72,74]. Also, it might be reasonable to limit access to the monitoring results: if individuals who are processed by the system have full access to their fairness analysis, they could use this to 'game' the system, i.e., they could use the synthetic inputs to slightly modify their own input such that they receive a better outcome. While more transparency for the user is generally desirable, this has to be kept in mind to avoid introducing new unfairness on a meta-level.

Acknowledgements This work is partially funded by DFG grant 389792660 as part of TRR 248 – CPEC, by VolkswagenStiftung as part of grants AZ 98514, 98513 and 98512 – EIS, by the European Regional Development Fund and the Saarland within the scope of (To)CERTAIN, and as part of STORM_SAFE, an Interreg project supported by the North Sea Programme of the European Regional Development Fund.

References

1. Abbas, H., Fainekos, G.E., Sankaranarayanan, S., Ivancic, F., Gupta, A.: Probabilistic temporal logic falsification of cyber-physical systems. ACM Trans. Embed. Comput. Syst. **12**(2s), 95:1–95:30 (2013). https://doi.org/10.1145/2465787.2465797
2. Alves, W.M., Rossi, P.H.: Who should get what? fairness judgments of the distribution of earnings. American journal of Sociology **84**(3), 541–564 (1978)
3. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine Bias (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
4. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)
5. Artistotle: The Nicomachean Ethics. Oxford worlds classics, Oxford University Press, Oxford (1998), translation by W.D. Ross. Edition by John L. Ackrill, and James O. Urmson.
6. Artistotle: Politics. Oxford worlds classics, Oxford University Press, Oxford (1998), translation by Ernest Barker. Edition by R. F. Stalley.
7. Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. L. Rev. **104**, 671 (2016)
8. Bathaee, Y.: The artificial intelligence black box and the failure of intent and causation. Harv. JL & Tech. **31**, 889 (2017)
9. Baum, D., Baum, K., Gros, T.P., Wolf, V.: XAI Requirements in Smart Production Processes: A Case Study. In: World Conference on Explainable Artificial Intelligence. pp. 3–24. Springer (2023)
10. Baum, K., Mantel, S., Schmidt, E., Speith, T.: From responsibility to reason-giving explainable artificial intelligence. Philosophy & Technology **35**(1), 12

(2022). https://doi.org/10.1007/s13347-022-00510-w, https://doi.org/10.1007/s13347-022-00510-w
11. Biewer, S.: Software Doping – Theory and Detection. Phd thesis, Universität des Saarlandes (2023). https://doi.org/10.22028/D291-40364, http://dx.doi.org/10.22028/D291-40364
12. Biewer, S., Baum, K., Sterz, S., Hermanns, H., Hetmank, S., Langer, M., Lauber-Rönsberg, A., Lehr, F.: Software doping analysis for human oversight. Formal Methods in System Design (2024). https://doi.org/10.1007/s10703-024-00445-2, to appear; preprint available at https://arxiv.org/abs/2308.06186
13. Biewer, S., D'Argenio, P.R., Hermanns, H.: Doping tests for cyber-physical systems. ACM Trans. Model. Comput. Simul. **31**(3), 16:1–16:27 (2021). https://doi.org/10.1145/3449354, https://doi.org/10.1145/3449354
14. Biewer, S., Finkbeiner, B., Hermanns, H., Köhl, M.A., Schnitzer, Y., Schwenger, M.: On the road with RTLola. Int. J. Softw. Tools Technol. Transf. **25**(2), 205–218 (2023). https://doi.org/10.1007/s10009-022-00689-5, https://doi.org/10.1007/s10009-022-00689-5
15. Biewer, S., Hermanns, H.: On the detection of doped software by falsification. In: Johnsen, E.B., Wimmer, M. (eds.) Fundamental Approaches to Software Engineering - 25th International Conference, FASE 2022, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2022, Munich, Germany, April 2-7, 2022, Proceedings. Lecture Notes in Computer Science, vol. 13241, pp. 71–91. Springer (2022). https://doi.org/10.1007/978-3-030-99429-7_4, https://doi.org/10.1007/978-3-030-99429-7_4
16. Binns, R.: On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 514–524. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372864, https://doi.org/10.1145/3351095.3372864
17. Bloem, R., Chatterjee, K., Greimel, K., Henzinger, T.A., Hofferek, G., Jobstmann, B., Könighofer, B., Könighofer, R.: Synthesizing robust systems. Acta Informatica **51**(3-4), 193–220 (2014). https://doi.org/10.1007/s00236-013-0191-5, https://doi.org/10.1007/s00236-013-0191-5
18. Borgesius, F.J.Z.: Strengthening legal protection against discrimination by algorithms and artificial intelligence. The International Journal of Human Rights **24**(10), 1572–1593 (2020). https://doi.org/10.1080/13642987.2020.1743976, https://doi.org/10.1080/13642987.2020.1743976
19. Burke, L.: The Death and Life of an Admissions Algorithm (2020), https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd
20. Chazette, L., Brunotte, W., Speith, T.: Exploring explainability: A definition, a model, and a knowledge catalogue. In: 2021 IEEE 29th International Requirements Engineering Conference (RE). pp. 197–208 (2021). https://doi.org/10.1109/RE51729.2021.00025
21. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017). https://doi.org/10.1089/big.2016.0047, https://doi.org/10.1089/big.2016.0047
22. D'Argenio, P.R., Barthe, G., Biewer, S., Finkbeiner, B., Hermanns, H.: Is your software on dope? - formal analysis of surreptitiously "enhanced" programs. In: Yang, H. (ed.) Programming Languages and Systems - 26th European Symposium on Programming, ESOP 2017, Held as Part of the European Joint Conferences

on Theory and Practice of Software, ETAPS 2017, Uppsala, Sweden, April 22-29, 2017, Proceedings. Lecture Notes in Computer Science, vol. 10201, pp. 83–110. Springer (2017). https://doi.org/10.1007/978-3-662-54434-1_4, https://doi.org/10.1007/978-3-662-54434-1_4
23. Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. Science advances **4**(1), eaao5580 (2018)
24. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
25. Dworkin, R.: What is equality? part 2: Equality of resources. Philosophy & Public Affairs **10**(4), 283–345 (1981), http://www.jstor.org/stable/2265047
26. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Human Factors **37**(1), 32–64 (1995). https://doi.org/10.1518/001872095779049543
27. Endsley, M.R.: From here to autonomy: Lessons learned from human–automation research. Human Factors **59**(1), 5–27 (2017). https://doi.org/10.1177/0018720816681350, https://doi.org/10.1177/0018720816681350, pMID: 28146676
28. European Court of Justice: C-356/12 - glatzel ecli:eu:c:2014:350 (2014), https://curia.europa.eu/juris/liste.jsf?language=en&num=C-356/12
29. European Union: Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act), provisional version that has been adopted by the European Parliament on 13 March 2024 (2024), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf
30. Ferrer, X., Nuenen, T.v., Such, J.M., Coté, M., Criado, N.: Bias and discrimination in AI: A cross-disciplinary perspective. IEEE Technology and Society Magazine **40**(2), 72–80 (2021). https://doi.org/10.1109/MTS.2021.3056293
31. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. Commun. ACM **64**(4), 136–143 (mar 2021). https://doi.org/10.1145/3433949, https://doi.org/10.1145/3433949
32. Gunning, D.: Explainable artificial intelligence (XAI) (darpa-baa-16-53). Tech. rep., Arlington, VA, USA (2016)
33. Guryan, J., Charles, K.K.: Taste-based or statistical discrimination: The economics of discrimination returns to its roots. The Economic Journal **123**(572), F417–F432 (2013), http://www.jstor.org/stable/42919257
34. Hartmann, F.: Diskriminierung durch Antidiskriminierungsrecht? Möglichkeiten und Grenzen eines postkategorialen Diskriminierungsschutzes in der Europäischen Union. EuZA - Europäische Zeitschrift für Arbeitsrecht p. 24 (2006)
35. Heaven, W.D.: Predictive policing algorithms are racist. They need to be dismantled. (2020), https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/
36. High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
37. Hough, L.M., Oswald, F.L., Ployhart, R.E.: Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. International Journal of Selection and Assessment **9**(1-2), 152–194 (2001)
38. Ilvento, C.: Metric learning for individual fairness. arXiv preprint arXiv:1906.00250 (2019)

39. Jacovi, A., Marasović, A., Miller, T., Goldberg, Y.: Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 624–635 (2021)
40. Jewson, N., Mason, D.: Modes of discrimination in the recruitment process: formalisation, fairness and efficiency. Sociology **20**(1), 43–63 (1986)
41. John, P.G., Vijaykeerthy, D., Saha, D.: Verifying individual fairness in machine learning models. In: Adams, R.P., Gogate, V. (eds.) Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020. Proceedings of Machine Learning Research, vol. 124, pp. 749–758. AUAI Press (2020), http://proceedings.mlr.press/v124/george-john20a.html
42. Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., Sterz, S.: On the relation of trust and explainability: Why to engineer for trustworthiness. In: Yue, T., Mirakhorli, M. (eds.) 29th IEEE International Requirements Engineering Conference Workshops, RE 2021 Workshops, Notre Dame, IN, USA, September 20-24, 2021. pp. 169–175. IEEE (2021). https://doi.org/10.1109/REW53955.2021.00031, https://doi.org/10.1109/REW53955.2021.00031
43. Lai, V., Tan, C.: On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 29–38 (2019)
44. Langer, M., Baum, K., Hartmann, K., Hessel, S., Speith, T., Wahl, J.: Explainability auditing for intelligent systems: A rationale for multi-disciplinary perspectives. In: Yue, T., Mirakhorli, M. (eds.) 29th IEEE International Requirements Engineering Conference Workshops, RE 2021 Workshops, Notre Dame, IN, USA, September 20-24, 2021. pp. 164–168. IEEE (2021). https://doi.org/10.1109/REW53955.2021.00030, https://doi.org/10.1109/REW53955.2021.00030
45. Langer, M., Baum, K., Schlicker, N.: Effective human oversight of ai-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs (2023). https://doi.org/10.31234/osf.io/ke256
46. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., Baum, K.: What do we want from explainable artificial intelligence (XAI)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif. Intell. **296**, 103473 (2021). https://doi.org/10.1016/j.artint.2021.103473, https://doi.org/10.1016/j.artint.2021.103473
47. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How We Analyzed the COMPAS Recidivism Algorithm (2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
48. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human factors **46**(1), 50–80 (2004)
49. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: A review of machine learning interpretability methods. Entropy **23**(1) (2021). https://doi.org/10.3390/e23010018, https://www.mdpi.com/1099-4300/23/1/18
50. Matthias, A.: The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology **6**(3), 175–183 (2004). https://doi.org/10.1007/s10676-004-3422-1
51. Mecacci, G., de Sio, F.S.: Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. Ethics and Information Technology **22**(2), 103–115 (2020). https://doi.org/10.1007/s10676-019-09519-w
52. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)

53. Methnani, L., Aler Tubella, A., Dignum, V., Theodorou, A.: Let me take over: Variable autonomy for meaningful human control. Frontiers in Artificial Intelligence **4** (2021). https://doi.org/10.3389/frai.2021.737072, https://www.frontiersin.org/article/10.3389/frai.2021.737072
54. Meurrens, S.: The Increasing Role of AI in Visa Processing (2021), https://canadianimmigrant.ca/immigrate/immigration-law/the-increasing-role-of-ai-in-visa-processing
55. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: Mapping the debate. Big Data & Society **3**(2), 2053951716679679 (2016). https://doi.org/10.1177/2053951716679679, https://doi.org/10.1177/2053951716679679
56. Molnar, C., Casalicchio, G., Bischl, B.: Interpretable machine learning - A brief history, state-of-the-art and challenges. In: Koprinska, I., Kamp, M., Appice, A., Loglisci, C., Antonie, L., Zimmermann, A., Guidotti, R., Özgöbek, Ö., Ribeiro, R.P., Gavaldà, R., Gama, J., Adilova, L., Krishnamurthy, Y., Ferreira, P.M., Malerba, D., Medeiros, I., Ceci, M., Manco, G., Masciari, E., Ras, Z.W., Christen, P., Ntoutsi, E., Schubert, E., Zimek, A., Monreale, A., Biecek, P., Rinzivillo, S., Kille, B., Lommatzsch, A., Gulla, J.A. (eds.) ECML PKDD 2020 Workshops - Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14-18, 2020, Proceedings. Communications in Computer and Information Science, vol. 1323, pp. 417–431. Springer (2020). https://doi.org/10.1007/978-3-030-65965-3_28, https://doi.org/10.1007/978-3-030-65965-3_28
57. Mukherjee, D., Yurochkin, M., Banerjee, M., Sun, Y.: Two simple ways to learn individual fairness metrics from data. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 7097–7107. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/mukherjee20a.html
58. Noorman, M.: Computing and Moral Responsibility. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Spring 2020 edn. (2020)
59. Nunes, I., Jannach, D.: A systematic review and taxonomy of explanations in decision support and recommender systems. User Modeling and User-Adapted Interaction **27**(3), 393–444 (2017)
60. O'Neil, C.: How algorithms rule our working lives (2016), https://www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives, Online; accessed: 2023-06-23
61. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, USA (2016)
62. Orcale: AI in human resources: The time is now (2019), https://www.oracle.com/a/ocom/docs/applications/hcm/oracle-ai-in-hr-wp.pdf
63. Organisation for Economic Co-operation and Development (OECD): Artificial intelligence, machine learning and big data in finance: Opportunities, challenges and implications for policy makers. Tech. rep., [París] : (2021), https://www.oecd.org/finance/financial-markets/Artificial-intelligence-machine-learning-big-data-in-finance.pdf
64. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Comput. Surv. **55**(3) (feb 2022). https://doi.org/10.1145/3494672, https://doi.org/10.1145/3494672
65. Rawls, J.: Justice as fairness: Political not metaphysical. Philosophy & Public Affairs **14**(3), 223–251 (1985), http://www.jstor.org/stable/2265349

66. Rawls, J.: A theory of justice: Revised edition. Harvard university press (1999)
67. Rawls, J.: Justice as fairness: A restatement. Harvard University Press (2001)
68. Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. CoRR **abs/1606.05386** (2016), http://arxiv.org/abs/1606.05386
69. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778
70. Rowe, T.: Can a risk of harm itself be a harm? Analysis **81**(4), 694–701 (2022). https://doi.org/10.1093/analys/anab033
71. Sanneman, L., Shah, J.A.: A situation awareness-based framework for design and evaluation of explainable AI. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 94–110. Springer (2020)
72. Schlicker, N., Langer, M.: Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In: Mensch und Computer 2021, pp. 325–329 (2021)
73. Schlicker, N., Langer, M., Ötting, S.K., Baum, K., König, C.J., Wallach, D.: What to expect from opening up 'black boxes'? comparing perceptions of justice between human and automated agents. Comput. Hum. Behav. **122**, 106837 (2021). https://doi.org/10.1016/j.chb.2021.106837, https://doi.org/10.1016/j.chb.2021.106837
74. Schlicker, N., Uhde, A., Baum, K., Hirsch, M., Langer, M.: Calibrated trust as a result of accurate trustworthiness assessment – introducing the trustworthiness assessment model. PsyArXiv Preprints (2022). https://doi.org/10.31234/osf.io/qhwvx
75. Santoni de Sio, F., van den Hoven, J.: Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI **5** (2018). https://doi.org/10.3389/frobt.2018.00015, https://www.frontiersin.org/article/10.3389/frobt.2018.00015
76. Smith, E., Vogell, H.: How Your Shadow Credit Score Could Decide Whether You Get an Apartment (2021), https://www.propublica.org/article/how-your-shadow-credit-score-could-decide-whether-you-get-an-apartment, Online; accessed: 2023-06-23
77. Speith, T.: A review of taxonomies of explainable artificial intelligence (XAI) methods. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. p. 2239–2250. FAccT '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3531146.3534639, https://doi.org/10.1145/3531146.3534639
78. Sterz, S., Baum, K., Lauber-Rönsberg, A., Hermanns, H.: Towards perspicuity requirements. In: Yue, T., Mirakhorli, M. (eds.) 29th IEEE International Requirements Engineering Conference Workshops, RE 2021 Workshops, Notre Dame, IN, USA, September 20-24, 2021. pp. 159–163. IEEE (2021). https://doi.org/10.1109/REW53955.2021.00029, https://doi.org/10.1109/REW53955.2021.00029
79. Tabuada, P., Balkan, A., Caliskan, S.Y., Shoukry, Y., Majumdar, R.: Input-output robustness for discrete systems. In: Proceedings of the 12th International Conference on Embedded Software, EMSOFT 2012, part of the Eighth Embedded Systems Week, ESWeek 2012, Tampere, Finland, October 7-12, 2012. pp. 217–226. ACM (2012), http://doi.acm.org/10.1145/2380356.2380396

80. Talbert, M.: Moral Responsibility. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Winter 2019 edn. (2019)
81. Thüsing, G.: European Labour Law, § 3 Protection against discrimination. C.H. Beck (2013)
82. United Nations Educational, Scientific and Cultural Organization (UNESCO): Recommendation on the ethics of artificial intelligence (2021), https://unesdoc.unesco.org/ark:/48223/pf0000380455
83. Wachter, S., Mittelstadt, B., Russell, C.: Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. W. Va. L. Rev. **123**, 735 (2020). https://doi.org/10.2139/ssrn.3792772, http://dx.doi.org/10.2139/ssrn.3792772
84. Washington State: Certification of Enrollment: Engrossed Substitute Senate Bill 6280 ('Washington State Facial Recognition Law') (2020), https://lawfilesext.leg.wa.gov/biennium/2019-20/Pdf/Bills/Senate%20Passed%20Legislature/6280-S.PL.pdf?q=20210513071229
85. Waters, A., Miikkulainen, R.: Grade: Machine learning support for graduate admissions. AI Magazine **35**(1), 64 (Mar 2014). https://doi.org/10.1609/aimag.v35i1.2504, https://ojs.aaai.org/index.php/aimagazine/article/view/2504
86. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in ranking: A survey. CoRR **abs/2103.14000** (2021), https://arxiv.org/abs/2103.14000
87. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International conference on machine learning. pp. 325–333. PMLR (2013)
88. Ziegert, J.C., Hanges, P.J.: Employment discrimination: the role of implicit attitudes, motivation, and a climate for racial bias. Journal of applied psychology **90**(3), 553 (2005)