Cambridge Forum on AI: Law and Governance (2025), **1**, e27, 1–25
doi:10.1017/cfl.2025.8

Cambridge Forum

RESEARCH ARTICLE

The Fair Game: Auditing & debiasing AI algorithms over time

Debabrota Basu and Udvas Das

Équipe Scool, Inria, Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, Lille, France
Corresponding author: Debabrota Basu; Email: debabrota.basu@inria.fr

(Received 1 November 2024; revised 18 February 2025; accepted 16 March 2025)

Abstract

An emerging field of AI, namely Fair Machine Learning (ML), aims to quantify different types of bias (also known as unfairness) exhibited in the predictions of ML algorithms, and to design new algorithms to mitigate them. Often, the definitions of bias used in the literature are observational, i.e. they use the input and output of a pre-trained algorithm to quantify a bias under concern. In reality, these definitions are often conflicting in nature and can only be deployed if either the ground truth is known or only in retrospect after deploying the algorithm. Thus, there is a gap between what we want Fair ML to achieve and what it does in a dynamic social environment. Hence, we propose an alternative dynamic mechanism, "Fair Game", to assure fairness in the predictions of an ML algorithm and to adapt its predictions as the society interacts with the algorithm over time. "Fair Game" puts together an Auditor and a Debiasing algorithm in a loop around an ML algorithm. The "Fair Game" puts these two components in a loop by leveraging Reinforcement Learning (RL). RL algorithms interact with an environment to take decisions, which yields new observations (also known as data/feedback) from the environment and in turn, adapts future decisions. RL is already used in algorithms with pre-fixed long-term fairness goals. "Fair Game" provides a unique framework where the fairness goals can be adapted over time by only modifying the auditor and the different biases it quantifies. Thus, "Fair Game" aims to simulate the evolution of ethical and legal frameworks in the society by creating an auditor which sends feedback to a debiasing algorithm deployed around an ML system. This allows us to develop a flexible and adaptive-over-time framework to build Fair ML systems pre- and post-deployment.

1. Introduction

In today's era, learning machines are at the epicentre of technological, social, economic and political developments, **their continuous evaluation and alignment are critical concerns**. In 2016, World Economic Forum[1] recognised the study of learning machines, i.e. machine learning (ML)(Barocas et al., 2023), and its superset artificial intelligence (AI) to be the driving force of the fourth industrial revolution.[2] Reckoning of modern AI not only motivates development of efficient learning algorithms to solve real-life problems but also aspires socially aligned deployment of them. This aspiration has pioneered the theoretical and algorithmic developments leading to *ethical*, *fair*, *robust* and *privacy-preserving AI*, in brief **responsible AI** (Dwork and Roth, 2014; Cheng et al., 2021; Liu et al., 2021; cas et al., 2023). The frontiers of responsible AI are well developed for static data distributions and

[1] https://www.weforum.org/meetings/world-economic-forum-annual-meeting-2016/
[2] https://www.weforum.org/agenda/2016/01/what-is-the-fourth-industrial-revolution

models, but their extensions to dynamic environments are limited to reinforcement learning (RL) with stationary dynamics (Sutton and Barto, 2018). Concurrently, the emerging trend of regulating AI poses novel regulations and quantifiers of risks induced by AI algorithms (Annas, 2003; Voigt and Von dem Bussche, 2017; Pardau, 2018; Madiega, 2021; Dabrowski and Suska, 2022). Followed by deployment of General Data Protection Regulation (GDPR)[3] in 2018 and upcoming EU AI act[4] in 2025–26, Europe pushes the frontiers of AI regulation and propels the paradigm of **algorithmic auditing**.[5] Specially, EU AI act[6] discusses like any publicly used technology AI should undergo an audit mechanism, where we aim to understand the impacts and limitations of using this technology, and why they are caused. **But these two approaches are presently unwed.**

Specifically, existing responsible AI algorithms fix a property (e.g. privacy, bias, robustness) first, then the theory is built to learn with these properties, and finally algorithms are designed to achieve optimal alignment (Dwork and Roth, 2014; Liu et al., 2021; cas et al., 2023). *This present approach leaves little room for the auditor feedback to be incorporated in AI algorithms except the broad design choices.*

A curious case: `AIRecruiter`.

Let us consider an AI algorithm that uses a dataset of resumes and successful recruitments to learn whether an applicant is worth recruiting or not by an organisation. Multiple platforms, such as Zoho Applicant Tracking System[7] and LinkedIn job platform,[8] are already used in practice. The designer would obviously want the `AIRecruiter` algorithm to be accurate, regulation-friendly and practically useful. To be regulation-friendly, `AIRecruiter` has to consider the questions of social alignment, such as privacy, bias and safety. In addition, the labour market and company's financial situations are dynamic. Thus, the recruitment policies and the drive to achieve social alignment of `AIRecruiter` evolve over time. `AIRecruiter` *demonstrates one of the many practical applications, where social alignment of an AI algorithm over time becomes imperative.*

Unbiasedness (fairness): Debiasing and auditing.

If `AIRecruiter` is trained on historical data with a dominant demography, it would typically be biased towards the "majority" and less generous to the minorities (Barocas et al., 2023). This is a common problem in many applications: gender bias against women in Amazon hiring system,[9] economical bias against students from poorer background in Scholastic Assessment Test (SAT) score-based college admission (Kidder, 2001), racial bias against defendants of colour in the COMPAS crime recidivism prediction system (Bagaric et al., 2019), to name a few. *This issue invokes the question what is the fair or unbiased way of learning best sequence of decisions and asks for conjoining ethics and AI.*

Debiasing algorithms.

Due to the ambiguous notions of fairness in society, researchers have proposed eclectic metrics for fairness (also called, unbiasedness) in offline and online settings (Kleine Buening et al., 2022; cas et al., 2023). Additionally, multiple algorithms are proposed to mitigate bias in predictions (Hort

[3] https://gdpr-info.eu/

[4] https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[5] https://auditing-fairness-tutorial.github.io/

[6] https://artificialintelligenceact.eu/

[7] https://www.zoho.com/fr/recruit/

[8] https://www.linkedin.com/business/talent/blog/applicant-tracking-system

[9] https://heinonline.org/HOL/LandingPage?handle=hein.journals/cdozo44&div=7&id=&page=

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

et al., 2024), which are mostly tailored for a given fairness metric. A few recent frameworks are proposed to analyse some group fairness metrics unitedly (Chzhen et al., 2020; Mangold et al., 2023; Ghosh et al., 2023a) but *it is not clear how to leverage them for a generic and dynamic bias mitigating algorithm design.*

Auditing algorithms.

The existing bias auditing algorithms follow two philosophies:[10] *verification* and *estimation*. Verification-based auditors aim to check whether an algorithm achieves a bias level below a desired threshold while using as small number of the data as possible (Goldwasser et al., 2021; Mutreja and Shafer, 2023). On the other hand, estimation-based auditors aim to directly estimate the bias level in the predictions of the algorithm while also being sample-efficient (Bastani et al., 2019; Ghosh et al., 2021; 2022a; Yan and Zhang, 2022; Ajarra et al., 2025). Though initial auditors used to evaluate the bias over full input data (Galhotra et al., 2017; Bellamy et al., 2019), distributional auditors have been developed to estimate bias over whole input data distribution (Yan and Zhang, 2022; Ajarra et al., 2025). Along with researchers in the community, we have developed multiple distributional auditors yielding probably approximately correct (PAC) estimations of different bias metrics while looking into a small number of samples (Figure 3). Recently, Ajarra et al., 2025 have derived lower bounds on number of samples required to estimate different metrics to propose a Fourier transform-based auditor that estimates all of the fairness metrics simultaneously. Authors also show that the auditor achieves constant manipulation proofness while scaling better than existing algorithms. This affirmatively concludes a quest for a universal and optimal bias auditors for offline algorithms.[11]

Need for a Fair Game: Adapting to dynamics of ethics and society.

Now, if a multi-national organisation uses the `AIRecruiter` over the years, the nature of acquired data from new job applicants suffers distribution shift over time due to world economics, labour market and other dynamics. Similarly, the regulations regarding bias also evolve from market to market and time to time. For example, ensuring gender equality in recruitment has been well argued since 1970s (Arrow, 1971), while ensuring demographic fairness through minority admissions is still under debate.

In present literature, auditors and debiasing algorithms are assumed to have static measures of bias oblivious to long-term dynamics and to be non-interactive over time. *As an application acquires new data over time, updates its model, and socially acceptable ethical norms and regulations also evolve, it motivates the vision of, and a continual alignment of AI with auditor-to-alignment loop.*

In this context, first, we propose the **Fair Game** framework (Section 3). Fair Game puts together an *Auditor* and a *Debiasing algorithm* in a loop around an ML algorithm and treats the long-term fairness as a game of an auditor that estimates a bias report and a debiasing algorithm that uses this bias report to rectify itself further. We propose an RL-based approach to realise this framework in practice.

Then, we specify a set of properties that Fair Game and its components should satisfy (Section 4).

(1) *Data frugal:* Data are the fuel of AI and often proprietary. Thus, it is hard to expect access to the complete dataset used by a large technology firm to audit its models and algorithms. Rather, often the regulations and auditing start by interacting with the software and understanding the bias induced by it. This follows access to a part of the dataset and models to audit it for rigorously. Thus, sample frugality is a fundamentally desired property of auditing algorithms.

[10] A detailed exposition is available at https://auditing-fairness-tutorial.github.io/.

[11] Another challenge for an auditor is achieving manipulation proofness (Yan and Zhang, 2022; Ajarra et al., 2025), which we are studying in **Regalia** project.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

This becomes even more prominent in dynamic settings as the datasets, models and measures of bias change over time.

(2) *Manipulation proof:* As auditing is a time-consuming and often legal mechanism, one has to consider the opportunity to update models up to a certain extent in order to match the speed of fast changing AI landscape. Thus, an ideal auditor should be able to adapt to minor shifts or manipulation in the data distributions, model properties, etc. So, we aim for building an auditor which satisfies constant manipulation proofness.

(3) *Adaptive and dynamic:* While designing an auditor for debiasing AI algorithm for real-life tasks, it should interact with the environment in a dynamic way, i.e. it should adapt to correct bias measures being in dynamic environment. For example, the notion of fairness can change over time, so an auditor should be dynamic to adapt to such properties that change over time. This is central conundrum to ensure long-term fairness.

(4) *Structured feedback:* In existing auditors, we study statistically efficient auditors yielding accurate global estimates of privacy leakage, bias and instability. All the ethical conundrums are not quantifiable from observed data but are subjective. Thus, auditors might provide preferential feedback on different sub-cases of predictions (Dai et al., 2023; Conitzer et al., 2024; Xiao et al., 2024; Yu et al., 2024). Additionally, an user who leverages an AI algorithm to make an informed decision may want to the response aligned with his/her/their preferences. An auditor can help an AI algorithm to tune their responses with respect to these preference alignments.

(5) *Stable equilibrium:* The other question is the existence of a stable equilibrium in Fair Game, which is a two-player game and both the players can manipulate the other. Under fixed bias measure but under dynamic shifts of data distribution and incremental model updates, Fair Game should be able to reach a stable equilibrium between the auditor and the AI algorithm.

Before proceeding to the contributions, in Section 2, we elaborate the algorithmic background for algorithmic auditing of bias, debiasing algorithms and their limitations as the basis of Fair Game. In Section 2.4, we briefly introduce concepts of RL as it is the algorithmic foundation of Fair Game.

2. Background: Static auditing and debiasing algorithms

2.1. Learning to predict from data: Fundamentals of ML models

An ML model (Mohri, 2018) is a function $f : \mathcal{X} \to \mathcal{Y}$ that maps a set of input features $\mathbf{X} \in \mathcal{X}$ to an output $Y \in \mathcal{Y}$.[12] For classifiers, the output space is a finite set of classes, i.e. $\{1, \dots, k\}$. For regressors, the output space is a d_o-dimensional space of real numbers. Training and deploying an ML model involves mainly *four* components: (a) training dataset, (b) model architecture, (c) loss function and (d) training algorithm.

Commonly, an ML model f is a parametric function, denoted as f_θ, with parameters $\theta \in \mathbb{R}^d$, and is trained on a **training dataset** $\mathbf{D}^T$, i.e. a *collection of n input–output pairs* $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ *generated from an underlying distribution* $\mathcal{D}$. The *exact parametric form of the ML model* is dictated by the choice of **model architecture**, which includes a wide variety of functions over last five decades. Training implies that given a model class $\mathcal{F} = \{f_\theta | \theta \in \Theta\}$, a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and training dataset $\mathbf{D}^T$, we aim to find the optimal parameter

$$\theta^* \triangleq \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} l(f_\theta(\mathbf{x}_i), y_i). \tag{1}$$

[12] **Notations:** We denote sets/vectors by **bold** letters, and the distributions by *calligraphic* letters. We express random variables in UPPERCASE, and an assignment of a random variable in lowercase.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

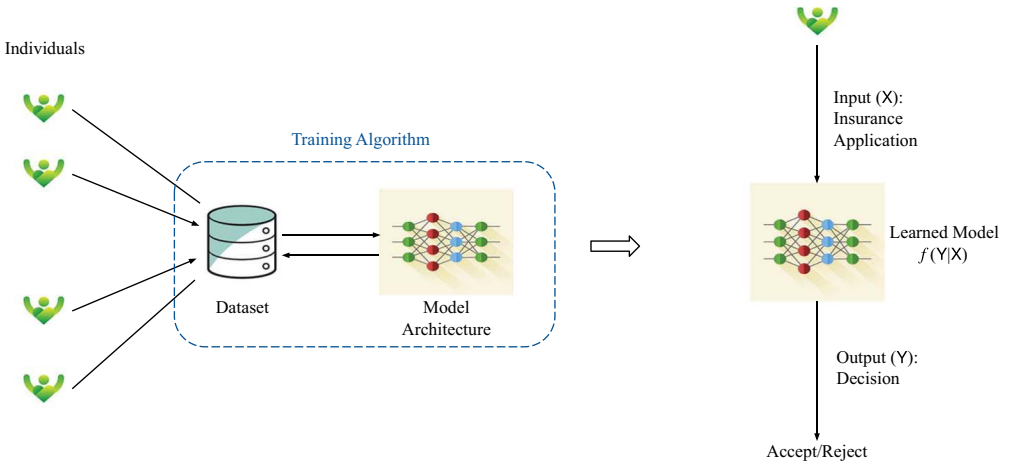

Figure 1. `AIRecruiter`: Training (left) and deploying (right) a machine learning model.

A **loss function** *measures badness of predictions made by the ML model with respect to the true output*. We commonly use cross entropy, i.e. $l(f_\theta(\mathbf{x}_i), y_i) \triangleq -y_i \log(f_\theta(\mathbf{x}_i))$, as the loss function for classification. For regression, we often use the square loss $l(f_\theta(\mathbf{x}_i), y_i) \triangleq (y_i - f_\theta(\mathbf{x}_i))^2$. Finally, *an optimisation algorithm is deployed to find the solution of the minimisation problem* in Equation (1). We refer to the optimiser as a **training algorithm**.

This procedure of training an ML model is called *empirical risk minimisation (ERM)* (Vapnik, 1991; Györfi et al., 2006; Feldman et al., 2012; Devroye et al., 2013). ERM is at the core of successfully training decision trees to large deep neural networks, like large language models (LLMs). The key statistical concept behind using ERM to train parametric ML models is that *if we have used large enough training dataset and the parametric family (aka model architecture) is expressive enough, the trained model can predict accurately (with high probability) for unseen input points coming from the same or close enough data generating distributions.* This property is called *generalisation ability* of an ML model and is often measured with its accuracy of predictions over a test dataset. A large part of statistical learning theory is dedicated to study this property for different types of data distributions, model architectures and training algorithms (Chaudhuri et al., 2011; Dandekar et al., 2018; Mohri, 2018; Tavara et al., 2021).

In Figure 1, we provide a schematic of this training and deployment schematic of ML models in the context of `AIRecruiter`. Specifically, `AIRecruiter` uses a historical dataset of resumes and future performance of job seekers to train a recruitment predicting ML model (*left* side of Figure 1). After successful training of the ML model, when it is deployed in practice, a resume of a candidate is sent through it and the model recommends accepting or rejecting the candidate (*right* side of Figure 1). This is well known as the binary classification problem. For example, Buening et al. (2022) showed similar mechanism and its nuances of gender and demographic bias in the context of college admissions. They use 15 years of data from Norwegian college admissions and examination performances to show that the historical data and fairness-oblivious ML models trained on it exhibit different types of bias under testing. We are aware of multiple such examples worldwide, such as racial bias in crime recidivism prediction in the COMPAS case (Angwin et al., 2016), gender bias in translating and completing phrases involving occupations by LLMs (Gorti et al., 2024), economic bias in SAT score-based college admissions (Kidder, 2001) to name a few.

However, discriminations in all these cases are prohibited up to different extents by laws of different countries (Blumrosen, 1967; Fiss, 1970; Madiega, 2021; Veale and Zuiderveen Borgesius, 2021) and

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

also are ethically unfair, in general (Novelli et al., 2023). This calls for design of bias auditing and debiasing ML algorithms.

2.2. Debiasing algorithms: State-of-the-art

A **debiasing algorithm** observes the input and output of an ML algorithm, the different quantifiers of bias estimated by the auditor and (if possible) the architecture of the ML algorithm to recalibrate the predictions of the ML algorithm such that the different quantifiers of bias are minimised (Lohia et al., 2019; Chouldechova and Roth, 2020; Mehrabi et al., 2021; Barocas et al., 2023; Caton and Haas, 2024).

Measures of bias.

Before proceeding to the debiasing algorithms, we provide a brief but formal introduction to different measures of bias.

In order to explain measures of bias, we consider a binary classification task (e.g. `AIRecruiter`) on a dataset $\mathbf{D}^T$ as a collection of triples $(\mathbf{X}, \mathbf{A}, Y)$ generated from an underlying distribution $\mathcal{D}$. $\mathbf{X} \triangleq \{X_1, \dots, X_{m_1}\}$ are non-sensitive features whereas $\mathbf{A} \triangleq \{A_1, \dots, A_{m_2}\}$ are categorical sensitive features. $Y \in \{0, 1\}$ is the binary label (or class) of $(\mathbf{X}, \mathbf{A})$. Each non-sensitive feature X_i is sampled from a continuous probability distribution $\mathcal{X}_i$, and each sensitive feature $A_j \in \{0, \dots, N_j\}$ is sampled from a discrete probability distribution $\mathcal{A}_j$. We use $(\mathbf{x}, \mathbf{a})$ to denote the feature-values of $(\mathbf{X}, \mathbf{A})$. For sensitive features, a valuation vector $\mathbf{a} = [a_1, .., a_{m_2}]$ is called a compound sensitive group. For example, consider $\mathbf{A} = \{\text{race, sex}\}$ where race $\in \{\text{Asian, Color, White}\}$ and sex $\in \{\text{female, male}\}$. Thus $\mathbf{a} = [\text{Asian, female}]$ is a compound sensitive group. We represent a binary classifier trained on the dataset $\mathbf{D}$ as $f : (\mathbf{X}, \mathbf{A}) \to \hat{Y}$. Here, $\hat{Y} \in \{0, 1\}$ is the predicted class of $(\mathbf{X}, \mathbf{A})$.

I. Measures of independence. **The prediction $\widehat{Y}$ of a classifier for an individual is independent of its sensitive feature** A. Mathematically, if $\widehat{Y}$ is binary variable, independence implies that for all $\mathbf{a}, \mathbf{b}$, $\Pr[\widehat{Y} = 1|\mathbf{A} = \mathbf{a}] = \Pr[\widehat{Y} = 1|\mathbf{A} = \mathbf{b}]$. Statistical/demographic parity (Kamishima et al., 2012; Zemel et al., 2013; Feldman et al., 2015; Corbett-Davies et al., 2017) measures deviation from independence of a classifier for a given data distribution.

Definition 1. (Statistical parity) $\text{SP} \triangleq \max_{\mathbf{a}} \Pr[\widehat{Y} = 1|\mathbf{A} = \mathbf{a}] - \min_{\mathbf{a}} \Pr[\widehat{Y} = 1|\mathbf{A} = \mathbf{a}]$.

Definition 2. (Demographic parity) $\text{DP} \triangleq \frac{\min_{\mathbf{a}} \Pr[\widehat{Y}=1|\mathbf{A}=\mathbf{a}]}{\max_{\mathbf{a}} \Pr[\widehat{Y}=1|\mathbf{A}=\mathbf{a}]}$.

The use of aforementioned metrics ensures equality of outcome across demographics. However, they can lead to accepting random people from majority and qualified people from minority, due to sample size disparity.

II. Measures of sufficiency. **The probability of positive outcome $\widehat{Y}$ for an individual given the true outcome is positive $Y = 1$ should be independent of its sensitive feature** A. Mathematically, if $\widehat{Y}$ and Y are binary variables, separation (equality of opportunity or equalised odds) implies that for all $\mathbf{a}, \mathbf{b}$, $\Pr[\widehat{Y} = 1|Y = 1, \mathbf{A} = \mathbf{a}] = \Pr[\widehat{Y} = 1|Y = 1, \mathbf{A} = \mathbf{b}]$. Separation metrics measure deviation from conditional independence of a classifier for a data distribution.

Definition 3. (Equalised odds) $\text{EO} \triangleq \max_{a,b} |\Pr[\widehat{Y} = 1|Y = 1, \mathbf{A} = \mathbf{a}] - \Pr[\widehat{Y} = 1|Y = 1, \mathbf{A} = \mathbf{b}]|$.

Incorporating EO (Hardt et al., 2016; Pleiss et al., 2017) as a measure of bias ensures equality of outcome for eligible individuals across demographics. But on the other hand, the true outcome Y is often not known in reality.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

III. Measures of calibration. **A classifier prediction $\widehat{Y}$ should be calibrated such that conditional probability of the true outcome $Y = 0/1$ should be independent of its sensitive feature A given the prediction being $\widehat{Y} = 0/1$.** Mathematically, if $\widehat{Y}$ and Y are binary variables, separation implies that for all a, b, $\Pr[Y = 1|\widehat{Y} = 1, \mathbf{A} = \mathbf{a}] = \Pr[Y = 1|\widehat{Y} = 1, \mathbf{A} = \mathbf{b}]$ and $\Pr[Y = 1|\widehat{Y} = 0, \mathbf{A} = \mathbf{a}] = \Pr[Y = 1|\widehat{Y} = 0, \mathbf{A} = \mathbf{b}]$.

Definition 4. (Predictive value parity (PVP))

$$\begin{aligned}\text{PVP} \triangleq \max\{ &\max_{\mathbf{a}} \Pr[Y = 1|\widehat{Y} = 1, \mathbf{A} = \mathbf{a}] - \min_{\mathbf{a}} \Pr[Y = 1|\widehat{Y} = 1, \mathbf{A} = \mathbf{a}], \\ &\max_{\mathbf{a}} \Pr[Y = 1|\widehat{Y} = 0, \mathbf{A} = \mathbf{a}] - \min_{\mathbf{a}} \Pr[Y = 1|\widehat{Y} = 0, \mathbf{A} = \mathbf{a}]\}.\end{aligned}$$

Such calibration measures like PVP (Hardt et al., 2016; Chouldechova, 2017) equalises chance of success given acceptance, but the acceptance largely depends on the choice of the classifier's utility function, which can be bias inducing.

There are other causal measures of bias than these three families of observational fairness metrics. We refer to Chouldechova (2017); Mehrabi et al. (2021) and Barocas et al. (2023) for further details on them.

Debiasing algorithms.

Given a measure of bias, there are three types of debiasing algorithms proposed in the literature: (a) pre-processing, (b) in-processing and (c) post-processing. These three families of algorithms intervene at three different parts of an ML model, i.e. training dataset, loss/training algorithm and final predictions post-deployment.

I. Pre-processing algorithms. These algorithms recognise that often the bias is induced from the historically biased data used for training the algorithm (Kidder, 2001; Bagaric et al., 2019; Barocas et al., 2023). When the data include a lot of samples for a demographic majority and very little for other minorities, then under ERM framework that tries to minimise the average loss to find the best parameters often lead to learning the patterns accurately for the majority and ignoring that of the minorities. Thus, pre-processing algorithms try to transform the training dataset and create a "repaired" dataset (Luong et al., 2011; Hajian and Domingo-Ferrer, 2012; Kamiran and Calders, 2012; Feldman et al., 2015; Heidari and Krause, 2018; Gordaliza et al., 2019; Salimi et al., 2019a; 2019b; 2019c). The advantage of pre-processing algorithms is that once the dataset is repaired, any model architecture and training algorithm can be used on top of it. The disadvantage is that it requires accessing and refurbishing the whole input dataset, which might include millions and billions of samples for large-scale deep neural networks. Thus, it becomes computationally intensive and requires retraining the downstream model for any update in bias measure.

II. In-processing algorithms. These algorithms aim to repair the fact that the classical ERM *tries to accurately learn on an average over the whole data distribution.* Thus, they either reweigh the input samples according to their sensitive features (Kamiran and Calders, 2012; Calders and Zliobaite, 2013; Jiang and Nachum, 2020) or modify the loss to maximise both fairness and accuracy (Agarwal et al., 2018; Celis et al., 2019; Chierichetti et al., 2019; Cotter et al., 2019). Some of the recent works have try and design optimisation algorithms that consider the fairness and accuracy simultaneously and iteratively during training. The advantage of these algorithms is that they achieve tightest fairness-accuracy trade-offs among the three families of debiasing algorithms. The disadvantage is that they require retraining an already established ML model from scratch, which is often time consuming, economically expensive and hard to convince the companies for whom models are central products.

III. Post-processing algorithms. The third family of debiasing algorithms approaches the problem post-training an ML model (Kleinberg et al., 2016; Chouldechova, 2017; Liu et al., 2017; 2019; Hébert-Johnson et al., 2018; Kim et al., 2018). They recognise the impact of bias of an ML model can be

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

stopped if only the predictions can be recalibrated according to their sensitive features and corresponding input–output distributions. Thus, post-processing approaches often apply transformations to model's output to improve fairness in predictions. If a debiasing algorithm can treat the ML model without accessing the data or training procedure, this is the only feasible approach to debias the ML model. Thus, it is the most flexible family of methods and can be used as a wrapper around existing algorithms. The only disadvantage is that we know it is not possible in one-shot to debias ML models with post-processing methods and might lead to sub-optimal accuracy levels in some cases.

For more details on the debiasing algorithms, we refer interested readers to detailed surveys and books published on this topic over years (Chouldechova and Roth, 2020; Mehrabi et al., 2021; Barocas et al., 2023; Caton and Haas, 2024).

Limitations.

We do not have a framework to accommodate the bias auditor feedback to improve debiasing of the algorithm under audit. This is fundamental to bridge the regulation-based and learning-based approaches to debiasing of an ML algorithm.

2.3. Auditors of bias: State-of-the-art

An **auditor** looks into the input and output pairs of an ML system and tries to measure different types of bias. Any publicly used technology presently undergoes an audit mechanism, where we aim to understand the impacts and limitations of using that technology, and why they are caused. In last decade, this has slowly but increasingly motivated development of statistically efficient auditors of bias, risk and privacy leakage caused by an ML algorithm. Though the initial auditors were specific to a training dataset and used the whole dataset to compute an estimate of bias (Bellamy et al., 2018; Pentyala et al., 2022), a need of sample-efficient estimation of bias has been felt. This has led to a PAC auditor that uses a fraction of data to produce the estimates of bias which are correct for the whole input–output data distributions (Albarghouthi et al., 2017; Bastani et al., 2019; Ghosh et al., 2021; Yan and Zhang, 2022; Ghosh et al., 2022c; Ajarra et al., 2025). Here, we formally define a PAC auditor of a distributional property of an ML model, such as different bias measures.

Definition 5. (PAC auditor (Ajarra et al., 2025)) Let $\mu : \mathbf{D}^T \times f_\theta \to \mathbb{R}$ be a computable[13] distributional property of an ML model f_θ. An algorithm $\mathcal{A}$ is a *PAC auditor* of property μ if for any $\epsilon, \delta \in (0,1)$, there exists a function $m(\epsilon,\delta)$ such that $\forall m \geq m(\epsilon,\delta)$ samples drawn from $\mathcal{D}$, it outputs an estimate $\widehat{\mu}_m$ satisfying

$$\mathbb{P}(|\widehat{\mu}_m - \mu| \leq \epsilon) \geq 1 - \delta\,. \tag{2}$$

In Figure 2, we provide a brief taxonomy of the existing bias auditing algorithms.

Components of a PAC auditor.

As illustrated in Figure 3, any PAC auditor consists of two components: (a) *sampler* and (b) *estimator*.

The **sampler** selects a bunch of input–output pairs from a dataset, which might or might not be same with the training dataset, and then send them further to query the ML model under audit. The information obtained by querying the algorithm depends on the access of the auditor. For example, an internal auditor of an organisation can obtain much detailed information, such as confidence

[13] Any distributional property of an ML model, including risk (Galindo and Tamayo, 2000; Paltrinieri et al., 2019; Assaf et al., 2020), individual fairness (Chouldechova and Roth, 2018; Pessach and Shmueli, 2022) and group fairness (Chouldechova and Roth, 2020), is computable given the existence of the mean estimators with uniformly random samples.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

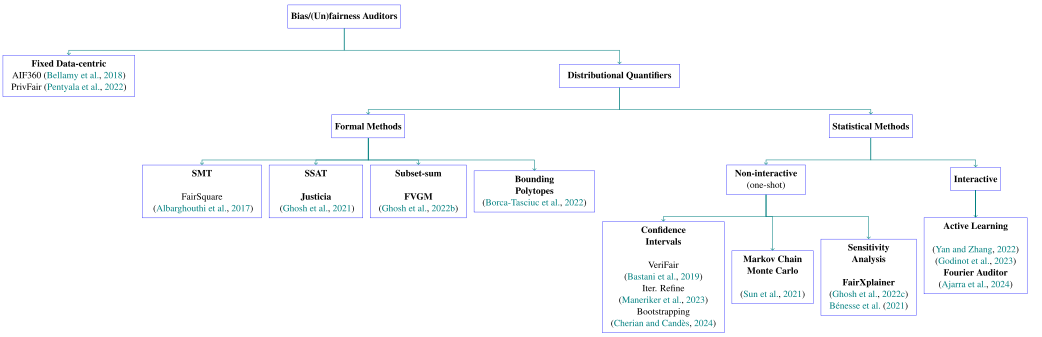

Figure 2. A taxonomy of bias auditors.

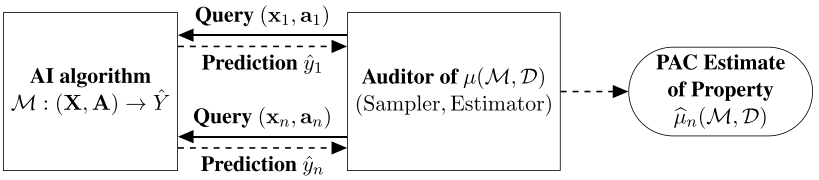

Figure 3. A generic schematic of a PAC auditor of a distributional property μ.

of predictions, gradients of loss at the query points, etc. This is called a white-box setting. On the other hand, an external auditor (e.g. a public body or a third-party company) might get only the predictions of the ML model for the queried samples. This is called a black-box setting. Initially, the auditing algorithms used only uniformly random samples from a dataset (e.g. the "Formal Methods" and "Non-interactive" auditors in Figure 2), but we know from statistics and learning theory that uniformly random sampling is less sample efficient than active sampling methods for most of the estimation problems. Specifically, the minimum number of samples required to PAC estimate mean of a distribution under uniform sampling is $\Omega(\epsilon^{-2} \ln(1/\delta))$, whereas the same for active sampling is $\Omega(\epsilon^{-1} \ln(1/\delta))$ (Yan and Zhang, 2022). Thus, to obtain an error of 1 per cent or below in PAC mean estimation, we get 100× decrease in the required number of samples. This has motivated design of active sampling mechanisms for auditing bias (Yan and Zhang, 2022; Godinot et al., 2023; Ajarra et al., 2025), and even other properties, like stability under different perturbations (Ajarra et al., 2025).

The **estimator** is the other fundamental component of an auditor. Typically, estimators try and quantify a specific measure of bias (Albarghouthi et al., 2017; Ghosh et al., 2023b; Cherian and Candès, 2024) or a family of bias measures (Bastani et al., 2019; Ghosh et al., 2021; 2022b; Ajarra et al., 2025). All of these estimators use the queried samples and their corresponding outputs to compute the bias measures. Designing efficient and stable estimators require understanding the structural properties of different bias measures properly and leveraging it in the algorithmic scheme. For example, Bastani et al. (2019) address bias estimation of programs as an Satisfiability Modulo Theory (SMT) problem, Ghosh et al. (2021) treat the same for any classifier as stochastic SAT (Boolean Satisfiability) problem, whereas Bastani et al. (2019) and Yan and Zhang (2022) bring it back to a set of conditional mean estimation problems. Ghosh et al. (2023b) and Bénesse et al. (2021) additionally aim for estimation of bias with attribution to features using global sensitivity analysis techniques from functional analysis. Ajarra et al. (2025) generalise this further by observing all the metrics of stability and bias are in the end impacts of different perturbations to the model distributions and can be computed using Fourier transformation of the input–output distribution of an ML model under audit.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

Limitations.
The present bias auditors, even the sample-efficiency wise optimal ones (Ajarra et al., 2025), can only audit static/offline AI algorithms accurately. The question to extend them for dynamic algorithms is still open. This is critical to create an auditor-to-alignment loop to audit and debias ML models over time.

2.4. Learning with feedback: A primer on RL

As we want to create a feedback mechanism between the auditor and debiasing algorithm and use their feedback to align ML models over time, we need to study the RL paradigm of ML that aims to learn about a dynamic environment while using only iterative feedback from the environment (Altman, 1999; Sutton and Barto, 2018). This ability of RL has been recognised, and thus it has been studied for long-term fairness and sequential decision-making problems, such as college admissions over years (Kleine Buening et al., 2022). The advantage is that "fair" RL algorithms allow us to refine biased decisions over time but like all existing debiasing algorithms they often need a fixed measure of bias and are tailor-made to optimise it (Gajane et al., 2022).

In RL (Sutton and Barto, 2018), a learning *agent* sequentially interacts with an *environment* by taking a sequence of *actions* and subsequently observing a sequence of *rewards* and changes in her *states*. Her goal is to compute a sequence of actions that yields as much reward as possible given a time limit. In other words, the agent aims to discover an optimal *policy*, i.e. an optimal mapping between her state and corresponding feasible actions leading to maximal accumulation of rewards. Two principal formulations of RL are bandits (Lattimore and Szepesvári, 2020) and Markov decision processes (MDPs) (Altman, 1999).

Bandit is an archetypal setting of RL with one state and a set of actions (Lattimore and Szepesvári, 2020). Each action corresponds to an unknown reward distribution. *The goal of the agent is to take a sequence of actions that both discover the optimal action and also allow maximal accumulation of reward.* The loss in accumulated rewards due to the unknown optimal action is called *regret*. Bandit algorithms are commonly designed with a theoretical analysis yielding an upper bound, i.e. a limit on the worst-case regret that it can incur (Basu et al., 2020; 2022; Azize and Basu, 2022; 2024; Azize et al., 2023). From information theory and statistics, we also know that this regret cannot be minimised more than a certain extent. This is called the *lower bound* on regret and indicates the fundamental hardness of the bandit. An algorithm is called **optimal** if its regret upper bound matches the lower bound up to constants.

In addition to bandits, **MDPs** include multiple states and a transition dynamics that dictates how taking an action transits the agent from one state to the other (Altman, 1999). An added challenge is to learn the transition dynamics and optimise the future rewards with it. The latter is called the *planning* problem. Thus, it requires generalising the optimal algorithm design tricks for bandits to handle the unknown transition dynamics and also to use an efficient optimiser to solve the planning problem. *Since RL considers the effects of sequential observations and learning with partial feedback from a dynamic environment* (Figure 4)*, it serves as the perfect paradigm to investigate the auditor-to-alignment over time.*

We specifically treat the Fair Game framework (elaborated in Section 3) as **performing RL in stochastic games.** On a positive note, success of many practical RL systems emerges in multi-agent settings, including playing games such as chess and Go (Silver et al., 2016; 2017), robotic manipulation with multiple connected arms (Gu et al., 2017), autonomous vehicle control in dynamic traffic and automated production facilities (Yang et al., 2020; Eriksson et al., 2022a; 2022b). Further advances in these problems critically depend on developing stable and agent incentive-compatible learning dynamics in multi-agent environment. Unfortunately, the mathematical framework upon which classical RL depends on is inadequate for multiagent learning, since it assumes an agent's environment is stationary and does not contain any adaptive agents. Classically, in multi-agent RL, these systems are

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

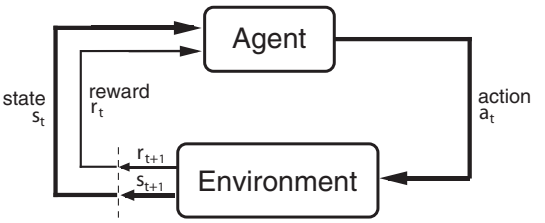

Figure 4. The feedback loop in reinforcement learning (RL).

treated as a stochastic game with a shared utility (Brown, 1951; Shapley, 1953). As an adaptive agent in this game only observes the outcomes of other's actions, the sequential and partial feedback emerges naturally. This connection has led to a growing line of works to understand limits of designing provably optimal RL algorithms for stochastic games (Giannou et al., 2022; Liu et al., 2022; Daskalakis et al., 2023). The existing analysis is often RL algorithm specific, i.e. they assume all the agents have agreed to play the same algorithm (Giannou et al., 2022). On the other hand, the lower bounds quantifying the statistical complexity of these problems are mostly available for either zero-sum games (Zhang et al., 2020; Fiegel et al., 2023) or large number of players (called mean-field games) (Elie et al., 2020).

3. Fair Game framework: Auditing & debiasing over time

Now, we formulate the Fair Game framework that aims to resolve two issues:

(1) incorporating the auditor feedback in the debiasing algorithm of an ML model,
(2) adapting to dynamics of society and ethical norm, and iteratively resolving the impacts of deploying ML models over time.

First, we provide a high-level overview of the framework and its components. Then, we further formulate the framework and the corresponding problem statement rigorously.

Fair Game: An overview.

In Figure 5, we illustrate the pipeline for an auditing to alignment feedback mechanism for any AI algorithm. First, an input dataset (Component (1)) is used to train an AI algorithm (Component (2)). This AI algorithm exhibits different alignment norms, also called model properties (Component (3)), such as privacy leakage, bias and instability (lack of robustness). An auditor (Component (4)) aims to accurately estimate the desired property (or properties) with minimal samples from a data pool, which might or might not match the input dataset depending on the degree of access available to the auditor. Then, the deployed auditor sends this feedback to the AI algorithm under audit, which is hardly used to incrementally debias the algorithm at present. Finally, we deploy another alignment algorithm (Component (5)) that leverages the feedback and other side information (e.g. preferences over outcomes), if available, to efficiently socially align the properties of the AI algorithm under audit.

In present literature, all of these components are assumed to be static over time. But as an application acquires new data over time, updates its model, and socially acceptable ethical norms and regulations also evolve, it motivates the conceptualisation of dynamic auditors. This also allows to bring in novel alignment properties from ethics and social sciences if they are computable or estimatable from observable data or their causal relations. This flexibility is essential as we still see plethora

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

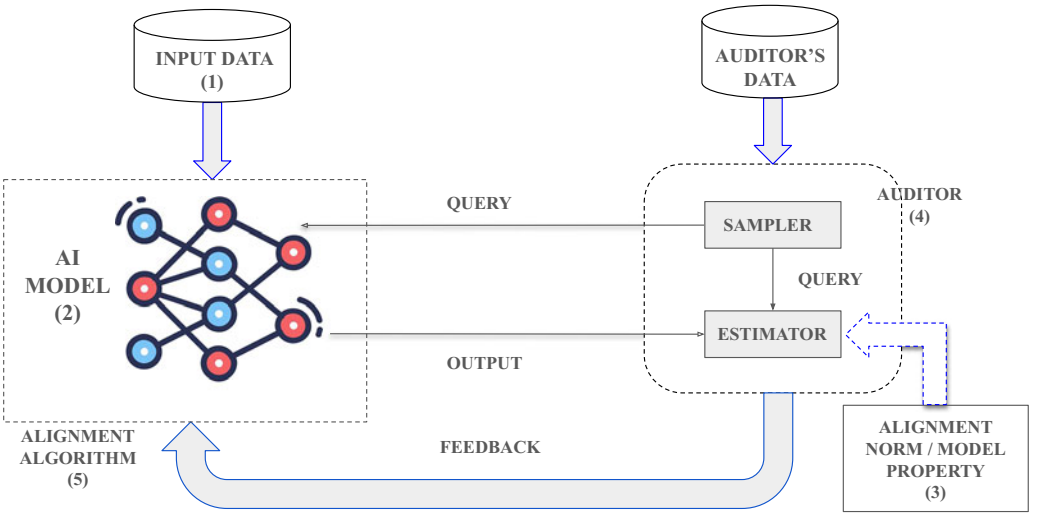

Figure 5. Components of an auditing to alignment mechanism for an AI algorithms.

of robustness and bias metrics to emerge after a decade of studying them, and this is a natural phenomenon as ethics evolve over time and we cannot know beforehand all the impacts of a young and blossoming technology like AI.

Fair Game: Mathematical formulation.

Let us consider the setting similar to Section 2.2, i.e. we have a binary classification model trained on a dataset $\mathbf{D}^T \triangleq \{(\mathbf{x}_i, \mathbf{a}_i, y_i)\}_{i=1}^m$, i.e. triplets of non-sensitive features, sensitive features and outputs generated from an underlying distribution $\mathcal{D}$. A model trained by minimising the average loss is denoted by f_{θ^*}. Given a measure of bias μ, f_{θ^*} exhibits a bias $\mu(f_{\theta^*}, \mathcal{D}) \geq 0$.

Now, let us consider that the underlying distribution changes with time $t \in \{1, 2, \ldots\}$. Thus, we denote the data distribution, the training dataset, the model and the property at time t as $\mathcal{D}_t$, $\mathbf{D}_t^T$, $f_{\theta^*,t}$ and μ_t, respectively. This structure defines the first three components in the Fair Game framework (Figure 5).

Under this dynamic setting, we first define **an anytime-accurate PAC auditor of bias** (Component (4)). The intuition is that an anytime-accurate PAC auditor of bias can achieve below ϵ error to estimate the desired bias measure as it evolves over time.

Definition 6. (Anytime-accurate PAC auditor) An auditor $\mathcal{A}$ is an *anytime-accurate PAC auditor* if for any $\epsilon, \delta \in (0, 1)$, there exists a function $m(\epsilon, \delta)$ such that $\forall m \geq m(\epsilon, \delta)$ samples drawn from $\mathcal{D}$, it outputs a mean estimate $\hat{\mu}_{m,t}$ at anytime t satisfying

$$\mathbb{P}(\forall t \in \{1, 2, \ldots\}, |\hat{\mu}_{m,t} - \mu_t| \leq \epsilon) \geq 1 - \delta\,. \tag{3}$$

Here, the probability is taken over all the stochastic dynamics of the data, the model and the auditing algorithm, if it uses randomised components.

Definition 6 means that with probability $1 - \delta$, an anytime-accurate PAC auditor yields an ϵ-accurate estimate of the property μ_t exhibited by the model at time t.

Now, we define the **dynamic debiasing algorithm**, which is the final component (Component (5)) required in Fair Game.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

Definition 7. (Dynamic debiasing algorithm) Given access to the data distribution, the training dataset, the model and an estimate of the property at time t, i.e. $\mathcal{D}_t$, $\mathbf{D}_t^T$, $f_{\theta^*,t}$ and $\hat{\mu}_t$, a dynamic debiasing algorithm $\mathcal{M}$ minimises the average bias over a given horizon $T \geq 1$, i.e.

$$V_T(\mathcal{M}) \triangleq \frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} \hat{\mu}_t(\mathcal{M}(f_{\theta^*,t}), \mathcal{M}(\mathcal{D}_t))\right]. \tag{4}$$

Here, the expectation is taken over all the stochastic dynamics of the data, the model, the bias estimate and the debiasing algorithm, if it uses randomised components.

As the average bias $V_T(\mathcal{M})$ of the dynamic debiasing algorithm tends to zero with increase in T, it implies that it is able to remove over time the bias in model predictions under the dynamic setup. In general, lower is the $V_T(\mathcal{M})$ better is the dynamic debiasing algorithm. From RL perspective, $V_T(\mathcal{M})$ is the value function measuring badness of the debiasing algorithm over time T, and the bias exhibited by it at time t, i.e. $\hat{\mu}_t(\mathcal{M}(f_{\theta^*,t}), \mathcal{M}(\mathcal{D}_t))$, is its cost function per-step.

Finally, with all these components, now we can formally define the Fair Game and its quantitative goals.

Definition 8. (Fair Game) Given access to the data distribution, the training dataset, the model and the property at any time t, i.e. $\mathcal{D}_t$, $\mathbf{D}_t^T$, $f_{\theta^*,t}$ and μ_t, the auditor-debiasing pair $(\mathcal{A}, \mathcal{M})$ plays a Fair Game by yielding anytime-accurate PAC estimates of the bias, i.e. $\{\mu, \hat{t}\}_{t=1}^T$, and minimising the average bias over time, i.e. $V_T(\mathcal{M})$, respectively.

We further define the **regret of the Fair Game** with an auditor-debiasing pair $(\mathcal{A}, \mathcal{M})$ as

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} \hat{\mu}_t(\mathcal{M}(f_{\theta^*,t}), \mathcal{M}(\mathcal{D}_t))\right] - \min_{\mathcal{A},\mathcal{M}} \frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} \mu_t(\mathcal{M}(f_{\theta^*,t}), \mathcal{M}(\mathcal{D}_t))\right]. \tag{5}$$

Regret of the Fair Game is the difference between the minimum bias achievable over time by any auditing and debiasing algorithm-pair and that achieved by a deployed system in practice for a given stream of datasets. We observe that lower regret of the Fair Game indicates higher efficiency of the auditor-debiasing pair $(\mathcal{A}, \mathcal{M})$. *Thus, given a dynamic dataset, an adaptive training algorithm and the evolving measures of bias, the goal of a Fair Game is to minimise its regret, while deploying anytime-accurate PAC auditors of bias and dynamic debiasing algorithms as two players with interactions.*

4. Challenges and opportunities to address the Fair Game

Now, we summarise the four desired properties of the Fair Game framework to audit and debias ML algorithms over time.

4.1. Data frugality and accuracy

The first pillar of the Fair Game is an anytime-accurate PAC auditor, but the auditor needs to query each of the updated models to conduct the estimation procedure. For external auditors and researchers designing algorithms, the data and access to proprietary ML models become the main bottleneck. Caton and Haas (2024) mention this dilemma as

> "This is a hard problem to solve: Companies cannot simply hand out data to researchers, and researchers cannot fix this problem on their own. There is a tension here between advancing the fairness state-of-the-art, privacy, and policy."

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

Thus, auditing over time brings us to the other pole of the AI world, where millions of datapoints are not available at all and we have to sharpen our statistical techniques to collect only informative data leading to accurate estimates. This leads to the first challenge in the Fair Game.

Challenge 1. Designing auditors that can use as minimum number of samples to yield as accurate estimate of bias as possible over dynamic data distributions and models.

Most sequential estimation and large-scale RL algorithms are known to be "data-greedy," which is the natural framework for auditing over time. This poses an opportunity to revisit the limits of statistical RL theory in the context of auditing over time as sample frugality becomes imperative.

4.2. Manipulation proof

Manipulation proof is an interesting and unique requirement of an auditor. Specially, an auditing mechanism is a top-down phenomenon in present AI technology scenario where ML models are changing in every economic quarter yielding more profit while we know little about their impacts in socioeconomic, cultural and personal lives. Manipulation proof auditing is specifically important due to two reasons.

(1) *Robustness to adversarial feedback.* To avoid being exposed or fined under auditing and the regulation hammer, a company can provide selected samples to the auditor which make them look fairer. This provides a partial view of the prediction distribution while not being too far from the true one (Yan and Zhang, 2022; Godinot et al., 2024).
(2) *Opportunity to evolve.* On the other hand, in the reckoning market of AI, a company might argue that they have to update their models "fast" to stay competitive. Thus, it is fair to give them an opportunity to change their models between two audits (Ajarra et al., 2025). This provides another motivation to design manipulation-proof auditors that can lead to easy acceptance of auditors in practice.

At this vantage point, we define PAC auditing with manipulation-proof certification that encompasses both the motivations.

Definition 9. (PAC auditing with manipulation-proof certification) For all $\epsilon, \delta \in (0,1)$, there exist a function $m : (0,1) \to \mathbb{N}$, such that for any probability measure $\mathcal{D} \in \mathcal{P}$, if S is a sample of size $m \geq m(\epsilon, \delta)$ sampled from $\mathcal{D}$, a PAC auditor with manipulation-proof certification yields

- **A correct estimate $\hat{\mu}_m$:**

$$\mathbb{P}\left[\sup_{h \in \mathcal{H}_\epsilon(f_\theta^*)} |\hat{\mu}_m(h) - \mu(h, \mathcal{D})| \geq \epsilon\right] \leq \delta.$$

- **A manipulation-proof region $\mathcal{H}_\epsilon(f_\theta^*)$:**

$$\mathbb{P}\left[\inf_{h \in \mathcal{H}_\epsilon^C(f_\theta^*)} |\hat{\mu}_m(h) - \mu(h, \mathcal{D})| \leq \epsilon\right] \leq \delta.$$

$\mathcal{H}_\epsilon(f_\theta^*)$ is a set of models with predictive distributions close to that of f_θ^*, and $\mathcal{H}_\epsilon^C(f_\theta^*)$ is its complement.

The manipulation-proof region $\mathcal{H}_\epsilon(f_\theta^*)$ allows the company under audit to change their models in regulated region around the present model. Manipulation-proofness aims to ensure whether

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

the AI model owner behaves adversarially and provides biased or obfuscated samples or asks for flexibility to update the model in-between two audits, the auditor should be able to estimate the bias robustly. An obfuscation of the samples or biasing them can be seen as a shift in the prediction distribution of the AI model under audit. Definition 9 claims that if we are sampling from any prediction distribution inside the manipulation-proof region $\mathcal{H}_\epsilon^C(f_\theta^*)$ around the true distribution under audit, the auditor can still yield good estimates of bias with high probability. In simple terms, the auditor is manipulation-proof in a regulated region around the true prediction distribution.

For the auditor, it provides an additional constraint, i.e. a region in which its bias estimate would vary minimally due to changes in predictions and input data. This goal is often in tension with accurate PAC estimation leading to a tension that we classically observe while designing any accurate by robust estimators (Huber, 1981).

This also makes PAC auditing with manipulation-proof certification a harder problem than PAC auditing. Yan and Zhang (2022) show an active learning-based procedure to achieve manipulation-proof auditing, while Godinot et al. (2024) show that manipulation-proof auditing can be harder as ML models gets larger and non-linear for a complex dataset. Ajarra et al. 2025 further show that if we use Fourier expansions of prediction distributions for auditing, we by default achieve manipulation-proofness with respect to changes in the smallest one-fourth coefficients. However, obtaining a universal complexity measure to quantify hardness of manipulation-proof auditing and comparing it with hardness of classical auditing still remains an open problem. At this point, auditing dynamic algorithms bring a stronger challenge.

Challenge 2. Designing manipulation-proof PAC auditors that can be accurate while computationally efficiently finds the manipulation-proof regions around evolving models over dynamic data distributions and models.

4.3. Adaptive and dynamic

In Section 3, we propose the formal framework of Fair Game. Specifically, Definition 8 formulates it rigorously as a two-player stochastic game. Thus, we propose to use the RL for stochastic game (Section 2.4) as the learning paradigm to resolve the Fair Game efficiently.

Specifically, last decade has seen a rise in responsible RL that aims to rigorously define and ensure privacy, unbiasedness and robustness along with utility over time. **Bias in RL** is studied as socio-political and economic policies (whether affirmative or punitive) interact with our society like the policies in RL do (e.g. `AIRecruiter`, college admissions over years, etc.). Thus, ensuring fairness in RL posits additional interesting and real-life problems (Kleine Buening et al., 2022). Researchers have studied effects of different fairness metrics on RL's performance and designed efficient algorithms to tackle them (Gajane et al., 2022), but the RL for two-player stochastic games still remains an open problem outside the worst-case and structure-oblivious scenarios. This brings us to the third challenge.

Challenge 3. Designing RL algorithms for two-player games with auditor-debiasing algorithm pairs deployed around evolving models over dynamic data distributions and models.

An opportunity arises from the growing study of RL and sequential estimation under constraints. Specifically, we note that *minimising bias over time with auditor feedback of bias is a special case of RL under constraints (Manerikar et al., 2023)*. Carlsson et al. (2024) and Das and Basu (2024) have derived lower bounds on performance that show how the optimal performance of an RL algorithm depends on the geometry of these constraints. Das

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

and Basu (2024) have reinforced the estimators constraint violations to achieve optimal performance, but these algorithms are still limited to the case of structure-oblivious and linear bandits. It is a scientific opportunity and challenge to extend them further to the Fair Game.

4.4. Structured and preferential feedback

With the growing real-life application of AI, it has become imperative to ensure their behaviour aligned with social norms and users' expectations. Especially, incorporating human preferential feedback in learning (or fine-tuning) process plays a vital role in aligning outputs from an LLM socially (Dai et al., 2023; Conitzer et al., 2024; Tao et al., 2024; Xiao et al., 2024). Reinforcement learning with human feedback (RLHF) enhances this alignment by using human judgments to fine-tune models, guiding them towards preferred actions and responses resulting in better model-adaptivity (Christiano et al., 2017; Ouyang et al., 2022; Song et al., 2024; Zhang et al., 2024), but RLHF can lead to over-optimisation for specific preferences, causing models to be overly specialised or biased, which repels adaptivity to diverse, unseen preferences in real-world applications (Christiano et al., 2017; Ziegler et al., 2019). In this context, Shukla and Basu (2024) propose the first preference-dependent lower bounds for bandits with multiple objectives and incomplete preferences. Even in this simpler setting, we observe that the preferences distort decision space. *But we understand very little how the incomplete preferences and high-dimensional features on continuous state-action spaces distort the decision space, which are closer to the LLMs.* Thus, the fourth challenge comes as follows.

Challenge 4. Designing debiasing algorithms that can optimally incorporate preferential and qualitative feedback of auditors to better debias the ML models over time.

4.5. Final destination: Existence of equilibrium?

The final question in any game is the existence of an **equilibrium**. We observe that bias measures can be PAC audited sample-efficiently with a universal auditor (Ajarra et al., 2025), while one can consider debiasing as a constraint optimisation problem of minimising loss while keeping the bias upper bounded. Thus, in presence of an auditor's feedback that accurately quantifies the bias and instability of a dynamic algorithm at any point of time, we can treat minimisation of bias and instability in a dynamic AI algorithm under constraints on the prediction distribution. This strategy has been studied in offline setting as ERM with distributional constraints. In addition, we know that the constraint violations can be used in feedback with a dynamic learning algorithm to achieve the desired safety and unbiasedness over time (Flet-Berliac and Basu, 2022). The generic framework to address them is to simulate a constraint-breaking adversary and a learner trying to avoid the adversary by only looking into its feedback. They use the same data stream to conduct their learning procedures. This poses the final two challenges.

Challenge 5. Can we achieve a stable equilibrium for the Fair Game when the measure of bias is fixed over time?

Challenge 6. Can we achieve a stable equilibrium for the Fair Game when the measure of bias is evolving over time?

We conjecture that the answer to the first challenge is affirmative while that of the second one depends on the changes in bias measures. Our intuition is based on the fact that bias measures are often conflicting and thus cannot be achieved simultaneously in a single game. One avenue to address

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

these problems will be to extend the growing literature on stochastic non-zero-sum games (Sorin, 1986; Zhang et al., 2020; Bai et al., 2022; Fiegel et al., 2023) with learners to the Fair Game setting, where utility of the auditor is to measure the constraint violation and that of the learner is to minimise loss in training while incorporating the auditor's feedback.

5. Bridging Fair Game and legal perspectives of audits

As Fair Game aims to bind the auditor and model owner into a single framework, it is a natural requirement to develop a legal framework for this and wonder how the present legal frameworks for auditing do or do not satisfy the requirements.

The landscape of law and algorithmic audits.

Le Merrer et al. 2023 provide an overview of the existing legal intricacies around algorithmic audits. We extend their insights to formalise the desired legal bindings for Fair Game.

The first legal concern for algorithmic auditing is data protection laws, such as the GDPR (Voigt and Von dem Bussche, 2017) in Europe. These laws safeguard user privacy and restrict data access to the algorithmic auditors (*Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases 2019*). In this context, Le Merrer et al. 2023 distinguish between two types of audits: Bobby Audits, which use real user data and thus face stronger legal scrutiny, and Sherlock Audits, which rely on synthetic data and are less legally restrictive but may have weaker evidential value. Fair Game naturally inherits these challenges of static auditing evoked by data protection laws. In this context, the rich literature of estimation and learning under privacy would play an instrumental role to develop efficient technical solutions (Chaudhuri et al., 2011; Dwork and Roth, 2014; Dandekar et al., 2018; Wang et al., 2020).

Another challenge in algorithmic auditing is due to intellectual property rights and trade secrets. Many companies argue that their algorithms are proprietary and, in turn, constraint the auditors access of information to examine them. The present law currently does not provide a unequivocal "right to audit." Thus, auditors often have to rely on indirect methods, such as scrapping the data from web services and digital platforms, or risk potential legal consequences, if they test proprietary algorithms without explicit permission. This propels the development of black-box auditors for estimating bias of AI models and is an active area of research (Ghosh et al., 2023b; Ajarra et al., 2025). This nuance is naturally covered by the Fair Game framework as it is applicable to both internal and external auditors with white-box and black-box access, respectively.

Liability is another challenge.[14] If auditors publicly disclose biases or discriminatory practices within an algorithm, they may face legal threats from companies seeking to protect their reputation and avoid regulatory penalties (Le Merrer et al., 2023). Additionally, how can we ensure that the companies would provide the auditors unbiased access to samples or update their models inside the prescribed manipulation-proof regions. Right now there is no reward or restriction on companies to not play adversarial during an audit. Thus, a legal binding of trust and liabilities between companies and auditors to create one single ecosystem is presently missing. Such developments have been seen for some other technologies, such as ISO[15] and IEEE[16] standards and certifications, and we need to develop one for emerging AI technologies. A unifying legal framework would be imperative to turn Fair Game into an effective paradigm for trustworthy deployment of AI.

[14]For example, Article 9 of French Civil Code states that "Each party has the burden of proving in accordance with the law the facts necessary for the success of its claim."

[15]https://www.iso.org/standards.html

[16]https://standards.ieee.org/ieee/24774/10126/

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

The legal environment for algorithmic auditing is still evolving. While current laws protect companies and user privacy, they must also facilitate responsible and ethical auditing. We now discuss about NYC Bias Audit Law, which is one of the real-life instance in this direction.

A bias auditing law in real-life: NYC Bias Audit Law.

As a real-life example of law enforcing algorithmic audits, we discuss the NYC Bias Audit Law or the Local Law 144 that has been mandated for algorithmic auditing of AI driven employment tools called "AEDT" (Automated Employment Decision Tool).[17] This is one of the first regulation enforced in the United States (specifically, New York City) that proposed auditing of digital platforms. Enacted in 2021 and enforced as of July 5, 2023, Local Law 144 is a significant step towards fairness and transparency in automated decision-making, particularly targeting employment and promotion processes.

Key components. 1. *Annual independent audits of bias.* Potential employers or organisations must conduct an independent bias audit of their AEDTs at least once per year. The audit evaluates bias in predictions across the protected demographics, specifically race, ethnicity and gender. The fairness metric used in auditing AEDTs is the ratio

$$\frac{\text{Selection rate for a category}}{\text{Selection rate of the most selected category}}.$$

This is a case-specific estimate of demographic parity (Definition 2) studied in fair ML literature. As the ratio goes to one, the AEDT is considered to be more fair.

2. *Candidate and employee notification.* Any employer must notify job applicants and employees at least 10 business days before using an AEDT in decision-making. It must include information about the tool's functionality and the applicant's right to request alternative evaluation methods. This aims to ensure transparency and privacy-compliant data processing in the AEDT tool.

3. *Public disclosure of audit results.* Any user of AEDTs must publicly post the results of their annual bias audits in their website demonstrating the system's fairness and impact on different protected demographics. This ensures the liability of AEDTs as a socially impactful technology as discussed before.

4. *Penalties for non-compliance.* Organisations failing to comply with these requirements face financial penalties, starting at 500 dollars for the first violation and increasing for subsequent infractions. This component aims to reinforce compliance of the user to the law and ensure unbiasedness in employment.

A failed attempt or a case for unified dynamic auditing? Despite of being a pioneering attempt towards bridging theoretical auditing and practical enforcement as a legal obligation for employers, the Local Law 144 has had many shortcomings.

(1) *Simplistic auditing process.* It is *too optimistic* that the auditor can use the final output and compute one of the many group fairness metrics to measure bias in order to create a comprehensive bias report. It does not promote for any causation of the bias and also no preference-based or case-to-case evaluation. It is very hard to develop any legal case with such simplistic audit. Fair Game brings in a richer framework to mitigate these issues.

(2) *Static auditing.* The auditing approach, specially the fairness metric suggested by the law, is not adaptive with time. It could neither leverage historical information in the auditing process nor can keep up with the changing notions of fairness. This is problematic as we cannot understand whether an AEDT tool is systematically biased or it is just one exceptional event. At this point, Fair Game brings in the dynamic perspective and proposes to treat the auditing and debiasing as a two player game, where both players have different incentives to play the game.

[17] AEDTs are real-life instances of our conceptual algorithm `AIRecruiter`.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

(3) *Lack of enforcing compliance.* Only one case of non-compliance was officially reported in six months after commencement though the authority was aware of hundreds of such cases. Also, there was no demand for third party auditor services, like proceptual.com (Janaro, 2023; Wright et al., 2024). The reasons are twofold. *First*, bureaucracy and lobbying delayed the enforcement of the law. Even after imposing it, NY city council loosened the laws and requirements and left several loops in the compliance rules within 24 hours of commencement. There is no purely technical solution to this. *Second*, the present law neither provides any reward or pressing reason for the AI model owner to comply with the auditing law nor it gives any flexibility and guideline to update the AI models between two audits. Fair Game proposes the dynamic and adaptive framework to overcome these issues.

6. Discussions and perspectives

With the growing use of AI algorithms for bouquets of real-life tasks such as autonomous recruitment, crime recidivism, autonomous college admission, etc., the need for debiasing these algorithms from any bias and aligning them with social norms has become crucial. Throughout this article, we have explained different sources of biases with real-life examples like AEDTs as well as the different approaches taken to mitigate them. We propose a dynamic mechanism "Fair Game" that assures to preserve fairness in prediction of ML algorithms that also adapt to changes as the society interacts with the algorithm over time. The "Fair Game" conjoins an *Auditor* and a *Debiasing Algorithm* in a loop leveraging the concepts of RL. Our approach addresses the lacunae of the existing approaches to debias an ML algorithm as they usually work well for static environments. "Fair Game" takes the process of debiasing to a dynamic environment that makes the process adaptive over time. We have also emphasised on some desirable properties of an auditor such as accuracy, data frugality, constant manipulation proofness, etc. Thus, "Fair Game" envisions an interactive approach to design statistically and computationally efficient continual auditors of AI algorithms and incorporate auditor's feedback in-a-loop to dynamically align these algorithms. This approach is novel, descriptive and more regulation-friendly and allows continual evaluation and alignment of AI algorithms.

Collaborative human and algorithmic auditing.

Finally, we would like to emphasise that the Fair Game does not aim to replace but aid the human AI auditors. This is necessary in order to connect the algorithmic audits with the available legal frameworks. Specifically, Fair Game enables the human auditors in two ways.

(i) For companies dealing with large number of user data, for example AIRecruiter, an auditor will have to manage and analyse a big and dynamic database. In these scenarios, a human-driven model becomes challenging in terms of resources and capacity. The AI auditor rather can provide statistically sound summaries of bias and other risks due to deployment of the AI algorithm that the human auditors and lawyers can use for further inspection.

(ii) In Fair Game, if an auditor is given a task to align an AI algorithm as per the requirements in Digital Marketing Act(DMA) or Digital Service Act (DSA) (for example, services provided by babl.ai), it demands intrinsic knowledge of the specific domain of application. This is where algorithmic auditors can enable us significantly by saving resources and creating a knowledge map that can be adapted over time.

But algorithms do not understand ethics till we design and tune them to it. In recent days, we have witnessed numerous proposals and research articles that advocate human intervention in aligning LLMs with social values in their decision-making whether by prompt engineering or leveraging RLHF (Wang et al., 2023; Ji et al., 2024; Shankar et al., 2024; Song et al., 2024). Thus, we envision the use of structural and preferential feedback (Section 4.4) in Fair Game, which allows human auditors

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

to intervene with case studies and preference over certain outcomes, and then the algorithmic auditor and the alignment algorithm further adapt to it.

Funding statement. We acknowledge the ANR JCJC project REPUBLIC (ANR-22-CE23-0003-01), the PEPR project FOUNDRY (ANR23-PEIA-0003) and the Regalia Project partnered by Inria and French Ministry of Finance.

Competing interests. This work and the authors are not funded by any private company or organisation.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning* (pp. 60–69). PMLR.

Ajarra, A., Ghosh, B., & Basu, D. (2025, April). Active Fourier auditor for estimating distributional properties of ML models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 15, pp. 15330–15338).

Albarghouthi, A., D'Antoni, L., Drews, S., & Nori, A. V. (2017). FairSquare: Probabilistic verification of program fairness. *Proceedings of the ACM on Programming Languages*, *1*(OOPSLA).

Altman, E. (1999). *Constrained Markov decision processes*. CRC Press, Vol. 7.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Annas, G. J. (2003). HIPAA regulations: A new era of medical-record privacy? *New England Journal of Medicine*, *348*, 1486.

Arrow, K. (1971). The theory of discrimination.

Assaf, D., Gutman, Y., Neuman, Y., Segal, G., Amit, S., Gefen-Halevi, S., Shilo, N., Epstein, A., Mor-Cohen, R., & Biber, A. (2020). Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Internal and Emergency medicine*, *15*, 1435–1443.

Azize, A., & Basu, D. (2022). When privacy meets partial information: A refined analysis of differentially private bandits. In *Advances in neural information processing systems* (1st ed., Vol. 22, pp.12). New Orleans, United States.

Azize, A., & Basu, D. (2024). Concentrated differential privacy for bandits. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (pp. 78–109). IEEE.

Azize, A., Jourdan, M., Marjani, A. A., & Basu, D. (2023). On the complexity of differentially private best-arm identification with fixed confidence. In *NeurIPS 2023 – Conference on Neural Information Processing Systems* (Vol. 36, pp. 71150–71194).

Bagaric, M., Hunter, D., & Stobbs, N. (2019). Erasing the bias against using artificial intelligence to predict future criminality: Algorithms are color blind and never tire. *University of Cincinnati Law Review*, *88*, 1037.

Bai, Y., Jin, C., Mei, S., & Yu, T. (2022). Near-optimal learning of extensive-form games with imperfect information. In *International Conference on Machine Learning* (pp. 1337–1382). PMLR.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bastani, O., Zhang, X., & Solar-Lezama, A. (2019). Probabilistic verification of fairness properties via concentration. *Proceedings of the ACM on Programming Languages*, *3*(OOPSLA), pp. 1–27.

Basu, D., Dimitrakakis, C., & Tossou, A. (2020). Privacy in multi-armed bandits: Fundamental definitions and lower bounds. *NeurIPS Workshop on Privacy Preserving Machine Learning*, 16.

Basu, D., Maillard, O.-A., & Mathieu, T. (2022). Bandits corrupted by nature: Lower bounds on regret and robust optimistic algorithm. Submitted to COLT'22.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., & Mojsilović, A. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4–1.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

Bénesse, C., Gamboa, F., Loubes, J.-M., & Boissin, T. (2021). Fairness seen as global sensitivity analysis.

Blumrosen, A. W. (1967). The duty of fair recruitment under the Civil Rights Act of 1964. *Rutgers Law Review*, *22*(1), 465.

Brown, G. W. (1951). Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, *13*(1), 374.

Buening, T. K., Segal, M., Basu, D., George, A.-M., & Dimitrakakis, C. (2022). On meritocracy in optimal set selection. In *EAAMO 2022 – Equity and access in algorithms, mechanisms, and optimization*. Arlington, United States, ACM.

Calders, T., & Zliobaite, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society: Data mining and profiling in large databases*. Springer, pp. 43–57.

Carlsson, E., Basu, D., Johansson, F. D., & Dubhashi, D. (2024). Pure exploration in bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics* (Vol. 238, pp. 334–342) of *Proceedings of Machine Learning Research (PMLR)*.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

cas, S., Hardt, M. and Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.
Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), 1–38.
Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 319–328).
Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, *12*(3), 1069–1109.
Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially responsible AI algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, *71*, 1137–1181.
Cherian, J. J., & Candès, E. J. (2024). Statistical inference for fairness auditing. *Journal of Machine Learning Research*, *25*(149), 1–49.
Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvtiskii, S. (2019). Matroids, matchings, and fairness. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 2212–2220). PMLR.
Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163.
Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *63*(5), 82–89. arXiv:1810.08810
Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, *63*(5), 82–89.
Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.
Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020). Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*, *33*, 7321–7331.
Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., et al. (2024). Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Forty-first International Conference on Machine Learning*, Vol. 235, (pp. 9346–9360).
Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806)
Cotter, A., Jiang, H., Gupta, M., Wang, S., Narayan, T., You, S., & Sridharan, K. (2019). Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, *20*(172), 1–59.
Dabrowski, L. D., & Suska, M. (2022). *The European Union digital single market: Europe's digital transformation*. Routledge.
Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., & Yang, Y. (2023). Safe RLHF: Safe reinforcement learning from human feedback. *ArXiv preprint arXiv:*2310.12773.
Dandekar, A., Basu, D., & Bressan, S. (2018). Differential privacy for regularised linear regression. In *International Conference on Database and Expert Systems Applications* (pp. 483–491). Springer.
Das, U., & Basu, D. (2024). Learning to explore with Lagrangians for bandits under unknown constraints. In *Seventeenth European Workshop on Reinforcement Learning*.
Daskalakis, C., Golowich, N., & Zhang, K. (2023). The complexity of Markov equilibrium in stochastic games. In *The Thirty Sixth Annual Conference on Learning Theory* (pp. 4180–4234). PMLR.
Devroye, L., Györfi, L., & Lugosi, G. (2013). *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 31.
Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4), 211–407.
Elie, R., Perolat, J., Laurière, M., Geist, M., & Pietquin, O. (2020). On the convergence of model free learning in mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 7143–7150).
Eriksson, H., Basu, D., Alibeigi, M., & Dimitrakakis, C. (2022a). Risk-sensitive Bayesian games for multi-agent reinforcement learning under policy uncertainty. In *OptLearnMAS@AAMAS*, Workshop on Optimization and Learning in Multiagent Systems at International Conference on Autonomous Agents and Multiagent Systems, Virtual, New Zealand.
Eriksson, H., Basu, D., Alibeigi, M., & Dimitrakakis, C. (2022b). SENTINEL: Taming uncertainty with ensemble-based distributional reinforcement learning. In *UAI 2022 – Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence of Proceedings of Machine Learning Research* (Vol. 180, pp. 631–640).
Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268).
Feldman, V., Guruswami, V., Raghavendra, P., & Wu, Y. (2012). Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, *41*(6), 1558–1590.
Fiegel, C., Ménard, P., Kozuno, T., Munos, R., Perchet, V., & Valko, M. (2023). Adapting to game trees in zero-sum imperfect information games. In *International Conference on Machine Learning* (pp. 10093–10135). PMLR.
Fiss, O. M. (1970). A theory of fair employment laws. *University of Chicago Law Review*, *38*, 235.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

Flet-Berliac, Y., & Basu, D. (2022). SAAC: Safe Reinforcement Learning as an Adversarial Game of Actor-Critics. In *RLDM 2022 – The Multi-disciplinary Conference on Reinforcement Learning and Decision Making*, Providence, United States. Accepted at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making.

Gajane, P., Saxena, A., Tavakol, M., Fletcher, G., & Pechenizkiy, M. (2022). Survey on fair reinforcement learning: Theory and practice. *ArXiv Preprint arXiv:*2205.10032.

Galhotra, S., Brun, Y., & Meliou, A. (2017). Fairness testing: Testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering* (pp. 498–510).

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, *15*(3), 107–143.

Ghosh, B., Basu, D., & Meel, K. (2023a). "How biased are your features?": Computing fairness influence functions with global sensitivity analysis. In *FAccT'23: The 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 138–148). ACM.

Ghosh, B., Basu, D., & Meel, K. S. (2021). Justicia: A stochastic sat approach to formally verify fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 7554–7563).

Ghosh, B., Basu, D., & Meel, K. S. (2022a). Algorithmic fairness verification with graphical models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 9539–9548).

Ghosh, B., Basu, D., & Meel, K. S. (2022b). Algorithmic fairness verification with graphical models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 9539–9548).

Ghosh, B., Basu, D., & Meel, K. S. (2022c, June). Algorithmic fairness verification with graphical models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 9, pp. 9539–9548).

Ghosh, B., Basu, D., & Meel, K. S. (2023b). "How biased are your features?": Computing fairness influence functions with global sensitivity analysis. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 138–148).

Giannou, A., Lotidis, K., Mertikopoulos, P., & Vlatakis-Gkaragkounis, E.-V. (2022). On the convergence of policy gradient methods to Nash equilibria in general stochastic games. *Advances in Neural Information Processing Systems*, *35*, 7128–7141.

Godinot, A., Le Merrer, E., Trédan, G., Penzo, C., & Taïani, F. (2023). Change-relaxed active fairness auditing. In *RJCIA 2023 - 21e Rencontres des Jeunes Chercheurs en Intelligence Artificiel*, CNIA. Strasbourg, France, Association Française pour l'Intelligence Artificielle, pp. 91–96.

Godinot, A., Le Merrer, E., Trédan, G., Penzo, C., & Taïani, F. (2024). Under manipulations, are some AI models harder to audit? In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (pp. 644–664). IEEE.

Goldwasser, S., Rothblum, G. N., Shafer, J., & Yehudayoff, A. (2021). Interactive proofs for verifying machine learning. In 12th Innovations in Theoretical Computer Science Conference (ITCS 2021) (Vol. 185, pp. 19). Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Gordaliza, P., Del Barrio, E., Fabrice, G., & Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning* (pp. 2357–2365). PMLR.

Gorti, A., Gaur, M., & Chadha, A. (2024). Unboxing occupational bias: Grounded debiasing LLMs with US labor data. In *Proceedings of the AAAI 2025 Symposium on AI Trustworthiness and Risk Assessment for Challenged Contexts (ATRACC)* (pp. 48–55). Washington, DC, USA: AAAI Press.

Gu, S., Holly, E., Lillicrap, T., & Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3389–3396). IEEE.

Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hajian, S., & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data engineering*, *25*(7), 1445–1459.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29.

Hébert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning* (pp. 1939–1948). PMLR.

Heidari, H., & Krause, A. (2018). Preventing disparate treatment in sequential decision making. In *IJCAI*, 2248–2254.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, *1*(2), 1–52.

Huber, P. J. (1981). Robust statistics. *Wiley Series in Probability and Mathematical Statistics*.

Janaro, C. (2023). NYC Local Law 144: A failed attempt at regulating AI in hiring.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., & Yang, Y. (2024). Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. *Advances in Neural Information Processing Systems*, *36*.

Jiang, H., & Nachum, O. (2020). Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics* (pp. 702–712). PMLR.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2012, Bristol, UK, September 24–28, 2012. Proceedings, Part II 23*. Springer, pp. 35–50.

Kidder, W. C. (2001). Does the LSAT mirror or magnify racial and ethnic differences in educational attainment: A study of equally achieving elite college students. *California Law Review*, *89*, 1055.

Kim, M., Reingold, O., & Rothblum, G. (2018). Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, 31.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Vol. 67, pp. 43:1–43:23). Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Kleine Buening, T., Segal, M., Basu, D., George, A.-M., & Dimitrakakis, C. (2022). On meritocracy in optimal set selection. In *Equity and access in algorithms, mechanisms, and optimization*. ACM, pp. 1–14.

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Le Merrer, E., Pons, R., & Trédan, G. (2023). Algorithmic audits of algorithms, and the law. *AI and Ethics*, *59*(2), 1–11.

Liu, J., Nogueira, M., Fernandes, J., & Kantarci, B. (2021). Adversarial machine learning: A multilayer review of the state-of-the-art and challenges for wireless and mobile systems. *IEEE Communications Surveys & Tutorials*, *24*(1), 123–159.

Liu, L. T., Simchowitz, M., & Hardt, M. (2019). The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning* (Vol. 97, pp. 4051–4060). PMLR.

Liu, Q., Szepesvári, C., & Jin, C. (2022). Sample-efficient reinforcement learning of partially observable Markov games. *Advances in Neural Information Processing Systems*, *35*, 18296–18308.

Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., & Parkes, D. C. (2017). Calibrated fairness in bandits. *arXiv preprint arXiv:*1707.01875.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2847–2851). IEEE.

Luong, B. T., Ruggieri, S., & Turini, F. (2011). K-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 502–510) San Diego, California, USA.

Madiega, T. (2021). Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 328.

Maneriker, P., Burley, C., & Parthasarathy, S. (2023). Online fairness auditing through iterative refinement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD'23. Association for Computing Machinery, pp. 1665–1676.

Mangold, P., Perrot, M., Bellet, A., & Tommasi, M. (2023). Differential privacy has bounded impact on fairness in classification. In *International Conference on Machine Learning* (pp. 23681–23705). PMLR.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning* (2nd ed.). MIT Press.

Mutreja, S., & Shafer, J. (2023). PAC verification of statistical algorithms. In *The Thirty Sixth Annual Conference on Learning Theory* (pp. 5021–5043). PMLR.

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). Taking AI risks seriously: A new assessment model for the AI act. *AI & SOCIETY*, 1–5.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing systems*, *35*, 27730–27744.

Paltrinieri, N., Comfort, L., & Reniers, G. (2019). Learning about risk: Machine learning for risk assessment. *Safety Science*, *118*, 475–486.

Pardau, S. L. (2018). The California Consumer Privacy Act: Towards a European-style privacy regime in the United States. *J. Tech. L. & Pol'y*, *23*, 68.

Pentyala, S., Neophytou, N., Nascimento, A., De Cock, M., & Farnadi, G. (2022). PrivFairFL: Privacy-preserving group fairness in federated learning. *ArXiv preprint arXiv:*2205.11584.

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, *55*(3), 1–44.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing systems*, 30.

Salimi, B., Howe, B., & Suciu, D. (2019a). Data management for causal algorithmic fairness. *ArXiv preprint arXiv:*1908.07924.

Salimi, B., Rodriguez, L., Howe, B., & Suciu, D. (2019b). Capuchin: Causal database repair for algorithmic fairness. *ArXiv preprint arXiv:*1902.08283.

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

Salimi, B., Rodriguez, L., Howe, B., & Suciu, D. (2019c). Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data* (19th ed., pp. 793–810). Amsterdam, Netherlands: ACM.

Shankar, S., Zamfirescu-Pereira, J., Hartmann, B., Parameswaran, A., & Arawjo, I. (2024). Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–14).

Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, *39*(10), 1095–1100

Shukla, A., & Basu, D. (2024). Preference-based pure exploration. In *Advances in Neural Information Processing Systems* (Vol. 38, p. 35). Vancouver (CA), Canada: NeurIPS.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., & Lanctot, M. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., & Graepel, T. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. 1, 19, arXiv preprint arXiv:1712.01815.

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., & Wang, H. (2024). Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, pp. 18990–18998).

Sorin, S. (1986). Asymptotic properties of a non-zero sum stochastic game. *International Journal of Game Theory*, *15*, 101–107.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed., pp. 352). MIT Press.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*(9), 346.

Tavara, S., Schliep, A., & Basu, D. (2021). In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 1, 459–467.

Vapnik, V. (1991). Principles of risk minimization for learning theory. *Advances in Neural Information Processing Systems*, *4*, 4.

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112.

Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR). *A Practical Guide, 1st Ed. Cham: Springer International Publishing*, *10*(3152676), 10–5555.

Wang, B., Gu, Q., Boedihardjo, M., Wang, L., Barekat, F., & Osher, S. J. (2020). DP-LSSGD: A stochastic optimization method to lift the utility in privacy-preserving ERM. In *Mathematical and Scientific Machine Learning* (Vol. 107, pp. 328–351). PMLR.

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., & Liu, Q. (2023). Aligning large language models with human: A survey. *ArXiv*,1 10619–10638.

Wright, L., Muenster, R. M., Vecchione, B., Qu, T., Cai, P., Smith, A., Investigators, C. S., Metcalf, J., Matias, J. N., et al. (2024). Null compliance: NYC Local Law 144 and the challenges of algorithm accountability. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1701–1713).

Xiao, J., Li, Z., Xie, X., Getzen, E., Fang, C., Long, Q., & Su, W. J. (2024). On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *ArXiv preprint arXiv:*2405.16455.

Yan, T., & Zhang, C. (2022, June). Active fairness auditing. In *International Conference on Machine Learning* (pp. 24929–24962). Maryland, USA: PMLR.

Yang, Y., Juntao, L., & Lingling, P. (2020). Multi-robot path planning based on a deep reinforcement learning DQN algorithm. *CAAI Transactions on Intelligence Technology*, *5*(3), 177–183.

Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., et al. (2024). RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vol. 18, pp. 13807–13816).

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In *International Conference on Machine Learning* (pp. 325–333). PMLR.

Zhang, K., Kakade, S., Basar, T., & Yang, L. (2020). Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, *33* (1), 1166–1178.

Zhang, T., Zeng, Z., Xiao, Y., Zhuang, H., Chen, C., Foulds, J., & Pan, S. (2024). GenderAlign: An alignment dataset for mitigating gender bias in large language models. *ArXiv Preprint arXiv:*2406.13925.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. *ArXiv Preprint arXiv:*1909.08593.

Debabrota Basu holds the Inria Starting Faculty Position in the Scool team (previously Sequel) of Inria Centre at Université de Lille and CNRS, France. He received his PhD in Computer Science from National University of Singapore (NUS). His research interest is to develop algorithms and analysis leading to theoretically grounded responsible AI systems. Specially, he

https://doi.org/10.1017/cfl.2025.8 Published online by Cambridge University Press

studies how to develop robust, private, fair and explainable algorithms for online learning, bandits and reinforcement learning problems. In 2022, ANR awarded the young researcher (JCJC) grant for his works on responsible AI. In 2024, he got elected as a scholar of European Learning and Intelligent Systems Society (ELLIS). His work on studying collective meritocracy in college admissions obtained Best Paper with Student Presenter Award at ACM EAAMO 2022. In IJCAI'23, he presented the tutorial on "Auditing Bias of Machine Learning Algorithms: Tools and Overview." For details, please visit: https://debabrota-basu.github.io/.

Udvas Das is a PhD student (2024–Present) in the Scool team (previously Sequel) of Inria Centre at Université de Lille and CNRS, France, under the supervision of Debabrota Basu. He graduated with Bachelor of Science degree in Statistics from St. Xavier's College, Kolkata (Autonomous) (2018–21). In 2023, he received a gold medal for his masters' degree in Computer Science from Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI). His research interest encompasses topics from statistics, machine learning, learning theory, reinforcement learning, information theory, fair machine learning, etc. His PhD focuses on studying the impact of constraints in learning under partial information setups, like bandits and RL. He received prestigious INSPIRE scholarship from Government of India. He presented his works in venues like EWRL'24, FoRLaC (ICML'24).

Cite this article: Basu D. and Das U. (2025). The Fair Game: Auditing & debiasing AI algorithms over time. *Cambridge Forum on AI: Law and Governance* **1**, e27, 1–25. https://doi.org/10.1017/cfl.2025.8