Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service"

Kornel Lewicki
Compliant & Accountable Systems Group
University of Cambridge
Cambridge, UK

Michelle Seng Ah Lee
Compliant & Accountable Systems Group
University of Cambridge
Cambridge, UK

Jennifer Cobbe
Compliant & Accountable Systems Group
University of Cambridge
Cambridge, UK

Jatinder Singh
Compliant & Accountable Systems Group
University of Cambridge
Cambridge, UK

ABSTRACT

"AI as a Service" (AIaaS) is a rapidly growing market, offering various plug-and-play AI services and tools. AIaaS enables its customers (users)—who may lack the expertise, data, and/or resources to develop their own systems—to easily build and integrate AI capabilities into their applications. Yet, it is known that AI systems can encapsulate biases and inequalities that can have societal impact. This paper argues that the context-sensitive nature of fairness is often incompatible with AIaaS' 'one-size-fits-all' approach, leading to issues and tensions. Specifically, we review and systematise the AIaaS space by proposing a taxonomy of AI services based on the levels of autonomy afforded to the user. We then critically examine the different categories of AIaaS, outlining how these services can lead to biases or be otherwise harmful in the context of end-user applications. In doing so, we seek to draw research attention to the challenges of this emerging area.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → **Computing / technology policy**; **Socio-technical systems**.

KEYWORDS

artificial intelligence, machine learning, bias, fairness, accountability, cloud, MLaaS, AIaaS, data-driven, algorithmic supply chains

ACM Reference Format:
Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2023. Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". In *CHI '23: ACM Human Factors in Computing, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3581463

CHI2023, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581463

1 INTRODUCTION

Artificial Intelligence is increasingly being offered "as a Service" (AIaaS). Marketed as "AI with no machine learning (ML) skills required" [7], AIaaS offers its users—primarily business customers—access to state-of-the-art AI capabilities, without the need for volumes of training data, expensive computational resources or lengthy development timescales that are generally required by traditional 'in-house' machine learning development.

At the same time, the reach and extent to which AI algorithms now pervade our daily lives has put AI under increased scrutiny. Various high-profile controversies (e.g. [10, 22, 112]) have raised concerns over the technology's potential to both perpetuate existing societal inequalities and introduce new types of discrimination [52, 67, 116]. As such, there are growing demands for greater levels of fairness, transparency and accountability over AI powered technologies (e.g., see [29, 45, 106, 153]).

In this context, the emergence of the AIaaS paradigm presents a number of challenges. First, given that these services aim to offer sophisticated AI capabilities at low cost and on demand to potentially anyone, often requiring little to no technical expertise, there is a risk that any algorithmic biases or other undesirable behaviours in the AIaaS system could be reproduced on *a far greater scale* across organisations [29, 72]. Further, with AIaaS products developed and sold by private for-profit companies, the inner workings of these commercial algorithms are often opaque or hidden (for various reasons), making it difficult for potential users to determine *if*, *when* and *why* bias manifests itself within these services.

Moreover, with an aim to appeal to the widest possible group of customers, many of these AI services are offered as generic 'AI building blocks', aiming to underpin a wide range of customers' applications in a variety of different contexts. However, as existing literature makes clear, *fairness is often contextual* [57, 90, 91, 131, 136]: a seemingly "fair" model designed for one context might be *misleading, inaccurate, or harmful* when applied to a different context [136]. This points to an inherent tension between the nuanced context-sensitive nature of fairness and the generic, 'one-size-fits-all' principle underpinning AIaaS.

In line with this, in this work, we treat the AIaaS paradigm as the object of study, considering whether enough thought has been put into the potential fairness risks associated with the development and adoption of AIaaS by an ever larger number of organisations. In particular, we question **whether the generic nature of such**

services *is* and *can be* appropriate given the diversity (i) of the potential applications in which they might be employed, and (ii) of the people, contexts, values, views and beliefs inherent to the real world.

Towards this, we first propose a taxonomy of AIaaS offerings based on three decreasing levels of autonomy and control afforded to the user: (1) *AutoML Platforms*, (2) *AI APIs*, and (3) *Fully-managed AI Services*. Through a combination of experimental evaluation, a consideration of topical examples, and theoretical analysis of a subset of AIaaS services from leading providers, we then critically examine each of the different categories of AIaaS offerings, outlining the novel fairness concerns and practical challenges arising from the development and use of these services in practice. Specifically, we consider the *perspective of the users* (customers) of these services (who can vary depending on the service type) and explore how the assumptions and constraints of different AIaaS models—which vary in levels of abstraction, transparency and control afforded by the provider to the user—can come into conflict with the contextual (and domain-specific) nature of fairness. This raises the risk of bias being embedded and propagating through the socio-technical systems supported by these services. Finally, we conclude by reflecting on what these concerns mean for the design, use and governance of AIaaS, identifying open challenges and suggesting potential paths forward.

In all, this paper (i) introduces and provides a taxonomy for describing various types of AIaaS services broken down by user perspectives so as to support an exploration of the concerns they raise; (ii) through the use of exemplars, elaborates the bias- and fairness-related issues and limitations of the 'one-size-fits-all' approach of AIaaS given the diverse range of contexts in which such services may be used; and (iii) discusses the wider impacts of the popularisation of AIaaS models and highlights multi-disciplinary research opportunities in this space. Our broader aim is to draw community attention to this nascent yet increasingly influential area and encourage discussions and ways forward regarding pragmatic approaches to the fair and responsible development and use of AI services.

2 BACKGROUND

To provide context for our exploration of fairness issues in AI services, we begin by briefly introducing AIaaS before describing some of the relevant emerging regulatory context, which both motivates the need for attention to be brought to issues of bias, and indicates some of the challenges that AIaaS brings to the space. We then highlight, with reference to related work, issues of algorithmic fairness and identify particular characteristics of AIaaS that can pose fairness-related challenges.

2.1 AI as a Service (AIaaS)

There is much interest in the adoption of AI among both business and governments alike [32, 44, 161]. However, the nature of AI development—requiring volumes of training data, specialised hardware for building and training models, and technical expertise in machine learning—presents a significant challenge for organisations wishing to integrate AI into their business processes [35, 44].

Recognising the opportunity, providers increasingly offer Artificial Intelligence solutions "as a service" (AIaaS) [29, 72]. AIaaS is advertised as a way of enabling developers without machine learning expertise to easily add AI capabilities to their applications. Various kinds of AI services exist (as we elaborate in §3), providing different functionality (e.g. video and image analysis, speech recognition and synthesis, etc.), with differing levels of customisation, and likely attracting different users (from experienced data scientists looking to accelerate their workflows to business users with no prior ML-knowledge). AIaaS promises its customers (the '**users**' of these services)[1] access to ML solutions without the need for volumes of training data, expensive computational resources or lengthy development times that are required by the traditional machine learning project development process. Indeed, the dominant AI service providers tend to be those very companies with significant access to the data, infrastructure and expertise required for effective ML development, and therefore are those best placed to develop and provide industry-leading AI services to derive (additional) revenue. On the other hand, the users of the services—who may lack the required resources, data, time or technical know-how to undertake and develop their own AI—gain direct access to AI-driven capabilities that may otherwise be beyond their reach [29].

Accordingly, and as was the case for the uptake of 'traditional' cloud [28], it is likely that AIaaS will become *the* primary means of AI proliferation and implementation across many industries and organisations (in contrast with 'in-house' AI development) [29, 130], given that such services reduce overheads and barriers to entry through providers offering and making accessible state-of-the-art tools/models. Indeed, we already see a consolidation of the AIaaS market predominantly around the few already dominant 'Big Tech' companies (primarily from two regions: U.S. and China), which further entrenches their position and extends their societal influence [4, 29]. This places AIaaS providers—primarily the cloud technology giants—in a powerful position. Crucially, AIaaS moves providers beyond merely offering supporting infrastructure for applications (as in cloud services) to directly enabling, facilitating, and underpinning customers' core applications and decision-making systems [29]. As such, there is a risk that through their services, AIaaS providers might propagate not only unintended bias, but also their own set of views, values and priorities. This can have wide-ranging social consequences given the vast numbers of users of such services and because of the potential issues that can arise from the diverse range of contexts in which these "generic" services can be used.

2.2 AI Regulation

With AI adoption growing rapidly in recent years, there has only recently been specific regulation targeting AI systems (though the use of AI is subject to general legal frameworks, such as data protection [29, 76] and non-discrimination [2]). There is, however, growing consensus that AI-specific regulation is needed to help address the undesirable and potentially harmful effects and ramifications of AI technologies. This often builds on non-legal sets of principles or ethical standards proposed by academics, civil society groups, and others for '*responsible*' and '*fair*' AI [36, 74, 105, 113, 135].

[1] In this paper we use the term 'users' to refer to customers of AI services.

In the US, for instance, a proposed *Algorithmic Accountability Act* would require impact assessments for AI systems—including in relation to fairness and bias concerns around automated decision-making systems—and seeks to address some transparency and accountability issues around AI [108, 147]. In the EU, a broader regulatory framework has been proposed in the form of the *Artificial Intelligence Act* (AI Act). Under the EU's proposed framework [30] systems which are deemed to pose an *unacceptable* risk (such as certain kinds of social scoring by public bodies) would be prohibited, while those responsible for systems classed as posing a *high* risk would be subject to risk management, transparency, accountability, and other requirements, responsibilities, and obligations. These include specific obligations to address concerns around fairness and bias, targeting both the datasets used to train models and the ongoing operation of systems. Those responsible for lower risk systems face more limited responsibilities, while systems in the lowest category of risk are effectively exempt.

While the increased legal and regulatory focus on fairness and bias concerns is welcome, neither the proposed Algorithmic Accountability Act in the US nor the EU's proposed AI Act properly account for the data-driven 'supply chain' [28, 139] dimension that AI services bring. The Algorithmic Accountability Act would apply to "covered entities" who "deploy" AI systems for "critical" decision-making (provided they meet certain other criteria, such as around revenue and size of customer base) and *also* to those who make systems for that purpose [147]. The AI Act will apply most of its requirements to "providers" of AI systems, who may in some cases be the users of AIaaS providers [30] (we later explore the fairness implications of this (§6.3)). Yet the divisions of responsibilities envisaged in these proposals do not address the fact that one provider will have many customers (users), each with their own contexts, use cases, and set of risks specific to their application. Moreover, these proposed laws do not address the information, skills, and power imbalances often present between AIaaS providers (again, typically large technology companies) and their service users (who may be much smaller organisations and/or lacking certain degrees of technical expertise), which may make it difficult if not impossible in practice for AIaaS users to identify and mitigate problems with systems that are not under their control. Nor do they address the more fundamental fact that such services are not under their control in the first place.

2.3 Algorithmic Fairness

As AI becomes increasingly pervasive, its propensity to both amplify existing biases and social inequities and create new ones has attracted considerable attention across a range of communities, including academics, policy-makers, industry, and civil society. Much of the initial work focused on developing quantitative definitions of fairness (see, e.g., [46, 61, 75, 93, 148]), and various technical methods for 'debiasing' AI models according to these mathematical formalisations (see, e.g., [3, 20, 50, 159]). *De-biasing* refers to the practice of removing undesired skews in the data and the model outcome, such as by equalising a metric of interest between groups. In academic literature, "unintended bias" is used to describe the different *sources* of harm that are introduced throughout an AI development lifecycle [89, 144]. A related taxonomy is the types of 'harm' as the consequences and scale of impact of AI [33]. While the former focuses on *how* the harm was introduced in the lifecycle, such as through poor data collection mechanisms, the latter describes *what type* of harm resulted from the unintended biases. While these taxonomies generalise on types of harms and their sources, the nuances of whether an algorithm is harmful and whether the potential risks of harm outweigh the potential benefits depend on the use case. More recently, aided by the interdisciplinary research in this area, there has been a growing realisation that *fairness is often contextual*; where considerations can differ across application domains, regions, cultures, and so forth, while some harmful algorithmic behaviours may only arise, or may only be recognised as harmful, when a system is used in a particular way, or in the presence of particular social or cultural dynamics [57, 90, 131, 136]. Accordingly, given the many complexities of fairness and its contextual nature, it is generally not possible to fully debias an AI system or to guarantee its fairness [80, 101, 122]; rather, the aim is to *mitigate* fairness-related harms and other unwanted consequences as much as possible [101, 136, 143].

A growing body of interdisciplinary work recognises that ML is part of a process [140], studying ML fairness through a sociotechnical lens which is aware of and actively considers the multitudes of social perspectives, actors, and interactions involved [136]. In the Human-Computer Interaction (HCI) domain, there has been work on studying public perceptions and expectations related to fairness in algorithmic systems. For instance, Binns et al. [16] and Woodruff et al. [155] identify gaps and dissonances in public understanding of this concept, while Srivastava et al. [141] found that these do not always align with existing mathematical definitions for fairness. Other work has focused on organisational challenges and barriers that practitioners face when attempting to build more responsible AI products and services [96, 97, 129], as well as considerations regarding fairness perceptions across cultures [131]. Research has also been directed specifically towards better understanding AI practitioners' needs and the development of frameworks, processes and tools to help assess and audit algorithmic systems for unfair, biased, or otherwise harmful behaviour (e.g., [14, 18, 97, 102, 128]) – both internally (within the organisations responsible for developing and maintaining these systems) and externally (by independent auditors, users and/or regulators). Particularly relevant here is the work exploring the issues and efficacy of fairness-specific tooling for supporting practitioners, for example, that of Holstein et al. [66], Lee et al. [91] and Deng et al. [40], which considered the perceptions and use of so-called "fairness toolkits" that aim to support ML practitioners with fairness concerns, finding significant disconnects between the tooling and the expectations, needs and practices of practitioners.

2.3.1 Fairness and AIaaS. Though issues of fairness pervade AI in general, there is the potential for AIaaS—in making widely-accessible, scalable and readily-available AI that is capable of underpinning a broad range of applications—to amplify AI fairness problems and introduce additional risks. Like any AI system, AI services can compound existing inequities by producing unfair outcomes, reinforcing pernicious stereotypes and disproportionately distributing negative consequences of technology to those already marginalised.

Each provider's AI services are offered to a range of possible customers, each of which can use the services for widely varying purposes. As such, issues of bias can arise both from the nature of the underlying ML model itself *and* the operating context in which a user employs that service. This means that AIaaS might not only contribute to replicating fairness issues at scale by propagating model's intrinsic biases (or other issues) to the vast number of user applications powered by a given service (see previous work including [22, 81, 133, 139]). Rather, it also raises challenges in terms of both identifying and mitigating fairness issues that can arise from a user's particular application of the service. This is because it can be difficult (if not impossible) to preempt all possible purposes for which a service might be used, while users employing these services may not have the knowledge, expertise, ability or access to assess the appropriateness of the service within their context and employ relevant safeguards. That is, as we discussed previously, one of the key difficulties in uncovering and mitigating unintended bias in AI systems is that many fairness issues will manifest only in particular contexts, meaning that understanding, let alone accounting for all bias concerns in an effort to create a 'generally applicable' AI service appears infeasible.

Importantly, much of the relevant literature on fairness, bias, and related harms has focused on AI systems built '*in-house*', where it is generally assumed that those producing the ML system are actively involved in building the model, and can adjust the model and/or the data it is trained on to mitigate fairness-related harms. There is some work on fairness relating to AI services, though this tends to focus on either exposing or 'auditing' certain AI services [22, 38, 126, 133], or devising tools, methods and interventions relevant for those building (rather than *using*) such services [12, 96, 106, 128]. However, there has been rather less focus on fairness issues and concerns in AI services as regards the perspective of those using the models and services of others. In arguing this as an area warranting more attention, in this paper we explore fairness from the perspectives of the users (customers) of AIaaS and highlight potential issues and challenges that arise from such services.

3 AIAAS TAXONOMY

To better understand and explore the fairness risks and challenges of "*AI as a Service*," it is important to first consider the overloaded nature of the term. We therefore now introduce a taxonomy to help characterise the various types of AI services currently on offer, in terms of their functionality and how they represent the different degrees of automation, complexity and user involvement. The taxonomy serves to provide a more precise description of the offerings and to better enable the exploration of the potential fairness risks and harms of different types of AIaaS.

Given AIaaS is an established business/marketing term, our guiding principle in structuring this space and deriving the taxonomy was to understand and reflect the ways in which the AIaaS providers distinguish these services themselves. Accordingly, as a first step, we have compiled a list of relevant AIaaS providers by consulting prior research [29, 37, 71, 133], market reports [130], blogs [25, 158] and news articles [60, 92]. Next, we comparatively reviewed the websites and marketing materials of AIaaS providers, elaborating the key characteristics, differences and similarities between the services offered, their naming, and the degree of user interaction and control. This process involved aligning the categorisation with the descriptions and terminology used in other AIaaS work [29, 71, 72, 157].

Table 1 represents the resulting taxonomy comprising three service categories: (1) *AutoML Platforms*, (2) *AI APIs*, and (3) *Fully-Managed AI services* (which we describe below). These categories reflect the perspective of users' (decreasing) level of involvement, required technical expertise and control over the service and the underlying ML model(s), as well as the increasingly specialised types of problems they address. These categories are not mutually exclusive in that some services may have characteristics that overlap the service types; however, identifying the service types helps indicate and highlight certain properties and characteristics that can have fairness implications in practice.

Consistent with other, more traditional cloud "*as a Service*" offerings (such as *SaaS* and *PaaS*), we focus AIaaS as referring to a set of *services* that are offered on demand, on a 'for a fee' basis and target (primarily) business customers. Here, we do not consider open-source and other tools that can be downloaded, used and operated directly (i.e. in a *non-service* context), such as AutoML frameworks (e.g. AutoKeras [73], TPOT [115]) and AI model libraries (e.g. Hugging Face [48], ModelZoo [107]), nor Large Language Models (e.g. BERT [42], GPT-3 [21]) or Generative AI tools (e.g. ChatGPT [117], DALL-E [118], Stable Diffusion [142] etc.), though we recognise that some of the points we discuss may also be applicable to these and other forms of AI tooling. Rather, we focus our discussion on offerings provided as commercial services on an on-demand basis, which raise considerations and implications relating to expectations around (i) levels of user expertise; (ii) the degree of visibility, control and access a user has; (iii) the potential for services to change and adapt (potentially without user knowledge); and (iv) the transactional ('per-use') nature of the service; and the legal and responsibility implications that come with commercial, data-driven, run-time supply chains involving several actors (§2.2).

Also, note that our aim here is not to form a comprehensive, future-proof, or definitive taxonomy of AIaaS, but rather represents a first attempt at categorising this evolving, market-driven space in a way that helps highlight the issues and brings the debate on fairness in AI to include AIaaS. As a living document, the taxonomy can be adapted and evolved as new considerations are brought forward and according to the particulars of the vendors and services, regulations, and so on. We now describe each of these categories in turn.

3.1 AutoML Platforms

Automated Machine Learning (AutoML for short) refers to the idea of building ML models with limited human intervention, either by automating certain stages in the ML workflow [78, 83, 87, 98] or by automating the entire ML development process [1, 6, 56, 104]. Here we focus on the latter, using "AutoML Platforms" to refer to services providing access to end-to-end, off-the-shelf AutoML functionality. In this way, by automating virtually every step of ML pipeline construction, AutoML Platforms allow users to create custom ML models (potentially) without ever seeing a single line of code used to create those models. From the user involvement

	AutoML Platforms	**AI APIs**	**Fully-managed AI Services**
Description	Automated model creation tools for building and deploying custom ML models.	Prebuilt AI models and services available via pay-per-use APIs.	AI models, services and platforms created and managed by an external third-party.
Use Cases	Predominantly tabular data classification and regression tasks.	Predominantly unstructured data problems such as object recognition, machine translation, text analytics etc.	Industry-specific problems such as candidate screening, recidivism prediction, healthcare allocation etc.
Control	User maintains control over the data used to train the model and some training configuration, while remaining steps of model creation are automated.	Models are already pretrained by the provider and available for use. The user has only limited or no control over the model.	Virtually entire control over the service is delegated to the service provider and the user interacts with the service via vendor provided "no code" UI.
Providers	Largest cloud providers and dedicated AutoML startups.	Predominantly largest cloud providers.	Dedicated tech companies and startups.
Level of technical expertise	Data scientists and developers with at least *some* AI knowledge.	Developers with technical but not necessarily AI knowledge.	Minimal; suitable for those with non-technical backgrounds (e.g. general business users).

Table 1: AIaaS Taxonomy including three types of AIaaS services and their characteristics.

perspective, the AutoML Platform model creation process is limited to broadly three steps: (1) provision of the dataset to be used for the model training, (2) selection of the feature to be predicted, and (3) specification of the training budget (the maximum time the AutoML service will spend training user's model). Likewise, the subsequent deployment of these models is often also limited to a simple click of a button, and the whole process is managed through a no-code UI.

In terms of providers, the current landscape of commercial AutoML Platforms is fast-growing and diverse. The commercial AutoML offerings can be categorised broadly into two groups: those offered by major cloud providers (e.g., Google, Microsoft), and those offered by dedicated AutoML organisations (often startups, e.g., DataRobot, H2O.ai, 4Paradigm) [157]. The former exist as part of larger ecosystems of the cloud providers' broader offerings and are tightly integrated within them, while the latter, being more standalone, provide end-to-end support through their own offerings or by facilitating integration with external tools and platforms [157].

3.2 AI APIs

AI APIs refer to a set of services offering access to a range of *pre-built* ML models that users can essentially 'plug' into their applications via pay-per-use third-party APIs. These services are offered on an 'as-is' basis and as such are available for instant use. Generally, to use a particular model, a user sends to the AIaaS-provider a request (typically by a webservices API) specifying the desired service functionality (e.g. detect faces, categorise text), and the data to be treated as inputs (e.g. an image or text file). After conducting the required authentication and authorisation checks (API keys, budget, etc.), the provider will process the request and return the ML model's response [71]. In this way, the AI APIs can be thought of as 'AI building blocks', where the customer uses one or more of such services, integrating them into a solution for their particular needs.

With respect to the services offered, we generally observe that the major AI APIs providers [7, 55, 68, 103] tend to focus on generic capabilities that are useful in a wide range of contexts, and broadly many offerings can be grouped into four key categories of service:

- *Data Analytics* services - enable the analysis of users's data and include capabilities for anomaly detection, content personalisation, product recommendation and so forth.
- *Text* services - offer AI capabilities in areas such as machine translation, text analytics, sentiment analysis, conversational interfaces etc.
- *Speech* services - comprise speech processing tools such as speech-to-text, text-to-speech, speech translation and speaker recognition.
- *Vision* services - support content identification and analysis within image and video data. Example services include capabilities in object recognition, facial analysis, facial recognition, video indexing etc.

3.3 Fully-managed AI Services

Recently, we can also observe a growing trend of AI capabilities being offered in the form of fully-managed AI services. Whereas the aforementioned AI API services predominantly focus on 'one-shot' functionalities (where a user sends over some input data and receives back the results of the ML model prediction for that particular data), here the offering tends to be more complex and focused, where the service seeks to supply a partial or complete business workflow. Fully-managed AI services might ingest a variety of customer data, consist of several models and steps before producing a final prediction outcome, come with dedicated UI platforms or hardware, and even interact directly with the customer's own users. These offerings typically offer a more specialised set of functionalities targeting niche industry-specific needs, examples including algorithmic hiring [65, 123], medical processes such as diagnosis and triage [70, 95, 124], and mental health self-care [62] to name but a few. Further, whereas the AI APIs generally need to be programmatically further integrated into users' own applications, fully managed services can also come as more complete products, providing dedicated mobile and web interfaces or desktop

applications and involve multiple layers of processing. As such, the fully managed AI services model resembles that of a traditional SaaS (Software-as-a-Service) model, which removes the user from the low-level, technical complexities of these systems.

4 METHODS

In this paper we elaborate the different AI service types, explaining some of their key characteristics from fairness perspective and some risks they entail. Our approach is based on a critical review of AIaaS vendors' claims and practices and empirical experimentation using select AIaaS services. For each service type, we outline a set of fairness risks and tensions that can occur as a result of (i) the decreasing levels of autonomy and transparency they afford to users, and at the same time (ii) increasing the level of abstraction away from the social context in which these systems will be deployed.

In outlining these fairness concerns and considerations, we draw from the existing body of work on algorithmic fairness, and highlight the ways in which these issues are enacted and potentially amplified through AIaaS. We support our argument with a set of technical experiments using select AI services from leading providers to explore these issues in-depth and concretely demonstrate how they can manifest themselves in practice. Specifically, using three real-world datasets commonly used in the Fair-ML literature (Adult [82], German Credit [79] and COMPAS [10]) we first explore how existing AutoML platforms—through their focus on optimising ML model performance according to a single, unconstrained, *fairness-unaware* objective—can inadvertently lead to developing models with poor levels of fairness. Second, we explore issues of bias in the context of AI APIs and demonstrate why without knowing *how* their services will be used, *where, when* and by *whom*, it is virtually impossible for these services to properly account for all possible concerns and issues that may be relevant. We further perform a qualitative analysis of providers' claims and practices (e.g. what they have disclosed about the performance of their systems, development and validation procedures, and bias mitigation support), as has been done in specific algorithmic auditing contexts [125], to examine the various other trade-offs providers need to consider and highlight particular causes for concern.

Importantly, the aim of our experiments is not to perform a comprehensive analysis of the bias issues in any particular AI service; and further, we stress that our exploration should be viewed as presenting simply a snapshot of current AIaaS landscape and some of its potential issues at a particular point in time (given these services may be updated or otherwise changed at any time). Rather, we use practical, tangible exemplars to expose some fundamental issues concerning bias and fairness for the different AIaaS categories, which we argue restrict the utility and appropriateness of these services given the diversity of the values, views and beliefs inherent in the real world.

5 AIAAS FAIRNESS CONCERNS

We now walk through and explore bias and fairness issues for the three AIaaS service types earlier described.

5.1 AutoML Platforms

We first consider the AIaaS model offering the greatest levels of user autonomy, which are the AutoML platforms. In the AutoML setting, the goal of the AI service is to create *custom* ML models in a fully-automated way. The user retains control over the training data used to build the model, but the remaining steps of the ML pipeline, from feature engineering to model selection and optimisation are automated (to varying degrees), with the user's involvement often limited to the selection of the feature to be predicted and the optimisation metric to be used when selecting the optimal model. On one hand, this approach allows users of all skill levels to quickly build state-of-the-art ML models; on the other hand, it also greatly limits their influence over model specifics. Further, with many commercial AutoML solutions tightly integrated within the service provider's infrastructure, user control over the created models is further restricted. We now discuss how these restrictions could cause tensions from the fairness perspective and constrain the ability for users to ensure fairness of the models built with such services.

5.1.1 Fairness-unaware optimisation. The fundamental premise of the AutoML paradigm is that the traditional 'manual' model creation process can be replaced by an automated creation and subsequent search over possible model architectures in order to find the '*optimal*' model. In the context of AutoML platforms, this 'optimality' usually takes the shape of the best predictive performance as measured by one of the user selected metrics such as *accuracy* or *log loss* – meaning that a model's fairness is typically not considered when choosing between the various models created in the AutoML process. However, for the many real-world applications where a model can impact people, it is not sufficient to only have high prediction accuracy. Indeed, there is an increasing expectation and demand that ML practitioners ensure that the outcomes of ML applications are fair in that they do not discriminate against certain groups or individuals [23, 44]. Unfortunately, as prior studies have shown [34, 120, 121, 156], prioritising such fairness-unaware model optimisation—e.g. by focusing solely on the model's accuracy—can lead to inadvertently prioritising models with higher levels of bias.

To demonstrate the issues that 'fairness-blind' optimisation can cause in the AutoML service context, we utilise Azure's AutoML service as an example service from a prominent provider (Microsoft), to build models for three real-world datasets commonly used in the fair-ML literature—Adult, German Credit and COMPAS [10, 79, 82]—and examine the resulting models' predictive accuracy and fairness. Figure 1 shows the fairness and performance of models created using the Azure AutoML service and highlights: (1) the model deemed by the platform as 'optimal' and (2) the model achieving best fairness according to the *demographic parity difference* metric [46]. We find that it is indeed often the case that as a result of selecting models solely by their predictive performance, the models deemed as 'optimal' might in fact exhibit more bias than possible alternatives (up to 12% difference for German dataset split on the sex attribute), while 'fairer' alternatives with only marginally lower levels of accuracy do exist. This is particularly important when one considers that not all AutoML platforms offer customers the opportunity to choose between the various models that were created as part of the AutoML model creation and search process. For instance, in case of Google's AutoML offering, the user is only provided with access

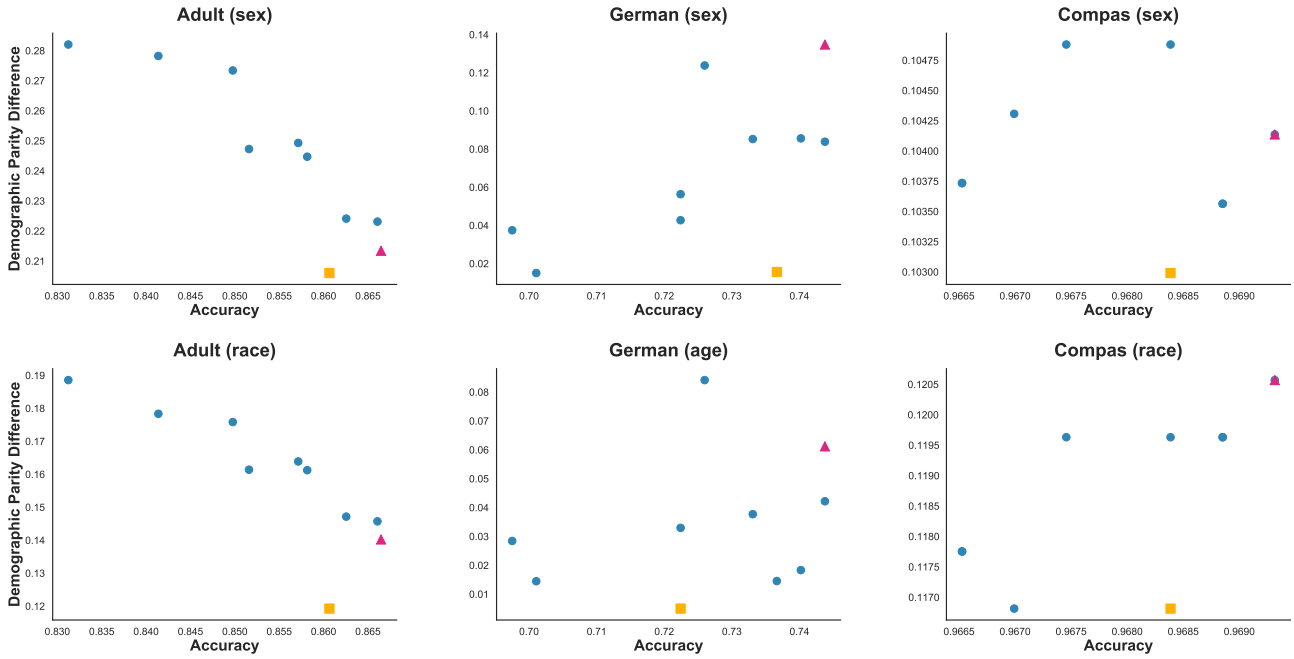

Figure 1: Fairness-accuracy trade-off of models created by Azure AutoML on Adult, German and COMPAS datasets measured across age, race and sex attributes. In pink (▲), the model selected by the algorithm as 'optimal' (higher is better); in orange (■), the model which achieves the best fairness according to "demographic parity difference" metric (lower is better). As a result of selecting models solely by their predictive performance, the models deemed as 'optimal' are often more biased than possible alternatives.

to the final model, deemed by the service as 'optimal', and does not have the opportunity to select other, potentially less-biased (or otherwise more appropriate), models that were created during the training process, to use and deploy.

The current reliance of AutoML frameworks on fairness-blind optimisation metrics combined with the inability to choose between the various models created in the AutoML process abstract away the complex, domain-specific notion of 'optimality' and its relational nature. In the face of the inevitable trade-offs between the real-world impact of changes in accuracy, fairness, privacy, consumer autonomy, and other considerations [89], such limitations restrict the users' ability to build a model that best reflects the values and objectives of their particular use case.

5.1.2 Constraints of proprietary platforms. The constraints of the proprietary model of commercial AutoML platforms extend beyond the model training process to the model parameter tuning, model selection, deployment and monitoring. This affects the scope of actions users could take to further interrogate created models or adjust them should any issues be observed.

Consider bias mitigation. The fair-ML literature has introduced various techniques that change the source data (pre-processing) or the outputs (post-processing) [50, 122, 160], as well as techniques that train a model to both maximise accuracy and increase fairness (in-processing) [34, 99], that could be used to mitigate potential bias issues we have outlined above. While various open source "fairness toolkits" have been developed to make these fairness methods more widely accessible to model developers [14, 18], currently these are largely unsupported by most AutoML platforms. Instead, AutoML customers' choice is generally limited to the provider's own, proprietary, bias mitigation functionality.

Yet again, in offering this functionality, AutoML providers make certain assumptions about how their services will be used and by whom, which may come into conflict with the customer's particular needs. For instance, IBM's AutoAI platform provides dedicated bias assessment and mitigation functionality in the form of its Watson OpenScale service allowing customers to uncover and mitigate bias according to the disparate impact fairness metric [69]. However, no further fairness metrics are supported. Here, using IBM's service, the user would have no choice in defining the fairness metric that is suitable for their particular use case, and is instead forced to use the sole fairness definition defined by the platform, which may be inappropriate for their application domain.

Note, however, that even if AutoML providers decided to make or extend the set of various fairness metrics and mitigation techniques available, these optimise for just narrow and particular definitions of fairness. Effectively tackling the fairness issues brought by real-world use cases requires more; indeed, addressing issues of fairness is a contextually specific exercise, requiring careful consideration of the domain specifics [91, 136, 149], of which fairness mitigation techniques and toolkits can be an important part of, but are not ends in themselves. As existing work has demonstrated, selecting

an appropriate fairness metric and mitigation method depending on the context of the use case is challenging even for experienced AI practitioners [91]. The AutoML paradigm of ML development will often (almost by definition) be unable to bring into account all the requisite factors, perspectives and constraints (which often require some domain expertise) that would otherwise need to be reconciled in addressing issues of fairness. Today's AutoML systems tend to limit customers' interaction with the system to a few specific decision points and present only limited to no information on how these systems work in operation and the complex processes behind the ML models generation and selection [88, 150–152, 157]. This 'blackbox' nature of AutoML operation results in a situation where users will often be unable to understand *how* and *why* AutoML systems make the choices they make [152], thus hindering their ability to effectively reason about and mitigate potential biases embedded in their outputs.

5.2 AI APIs

While AutoML services aim to automate the creation of *custom* machine learning models based on users' data and needs, AI APIs provide access to readily-available *pre-built* AI models that allow users to instantly integrate AI capabilities into their applications. This instant availability, however, generally comes at the cost of customisation. In the AI APIs setting, generally the user has no control over the dataset used to train the model or influence over the specifics of the model itself.[2] And while providers market these services with broad statements such as "infuse powerful AI capabilities into your apps and workflows" [68], "get quality and accuracy from continuously-learning APIs" [7] it is generally unclear what the performance of these services exactly is, how have they been trained, and how have they been evaluated. Below, we examine how the 'one-size-fits-all' ideal that underpins AI APIs, coupled with the inherent opacity of these services, can lead to potential bias and fairness issues and tensions when they are applied in the user-specific application contexts.

5.2.1 Universality. It is well acknowledged that machine learning models, such as those offered by the API providers, are prone to exhibiting various kinds of discriminatory behaviours, including different error rates across demographic groups [22, 126, 127], stereotyping [64, 163] or minority groups under-representation [134, 137]. Many of these works conclude that these issues are primarily a reflection of training set deficiencies, namely the under-representation of certain parts of the input space [15, 22] and the lack of awareness, consideration and attention of the people building these systems [97, 136, 146]. Consequently, a significant body of work argues for careful consideration of the social dimension in which the model will be applied, which requires the use of training data that is appropriate for capturing the full diversity of the context [27, 128, 144]. As we argue next, however, the nature of the AI API service provision model—which aims to provide *universal* 'AI building blocks' that can be used and deployed in a range of customer applications, without knowing the particulars of the customer's specific usage contexts—may render these services inherently incapable of addressing the representation issue.

Consider facial analysis services as an example, an area that has had some prior attention on issues of bias [22, 58, 126]. These services, generally available 'off-the-shelf' to anyone, aim to determine an individual's facial characteristics including physical or demographic traits based on an image of their face. However, human faces are not a homogeneous group [77]; the set of images used to train and evaluate the underlying model will have a significant influence over the model's behaviour and reported performance. Without knowing specifically how and where their customers will use their services, the AI API providers are at constant risk of failing to envision, let alone account for, all the various contexts their models' might encounter. Consequently, they risk their services failing to be representative of particular regions, cultures, groups or individuals.

To illustrate this difficulty in ensuring and evaluating a model's general representativeness, we examine the performance of the APIs from Amazon, Baidu, Face++, and Microsoft for age estimation tasks[3] across race groups, using a subset of UTKFace dataset [162]. First, we consider the APIs performance on unitary subgroups. Table 2 shows that all APIs perform relatively well, displaying no obvious differences between different race groups. However, following Buolamwini & Gebru's "Gender Shades" study and others [22, 126, 127]) and introducing *intersectional* considerations (across categories or characteristics)—which are said to provide a more complete picture of the biases that may exist in an AI system [22, 127]—reveals the existence of undesirable discrepancies. As shown in Table 3, disaggregating these subgroups into populations of under and over 60 years old reveals that all classifiers perform worse on the age estimation task for the 60+ population. Notably, Amazon's API age estimates on the 60+ subgroup exhibit an average error of almost 20 years for the Asian subgroup and over at least 15.5 years for all remaining groups. Further, as seen in Table 4, disaggregating those subgroups even further into UN's standard age brackets [114], makes this lack of representation of certain subgroups even more apparent. This is referred to as "representation bias" (see [89, 144]), or an unintended skewing of model outcomes due to the dataset's representativeness. The consequences are a "quality of service harm" [33], or the disparities in how well a model works for different groups of people, resulting in different experiences.

While some aspects of bias can potentially be reduced to *some* extent, in some situations, for example by training models on more diverse and representative datasets [126], this assumes the relevant contexts were uncovered and accounted for. With no easily discernible and limited finite set of 'protected classes' to rely on, there is no generally accepted limit to the number of potential categories bias can be evaluated on. Here we focused on the legally protected characteristic of age, but we might have equally well examined images of people with facial tattoos, brown eyes, glasses, headwear or other attributes, which could also reveal certain biases (see e.g. [41, 54]). In general, what the above shows is that understanding

[2]Note that there are exceptions, for example, services such as Amazon's Rekognition Custom Labels and Google's Vision AutoML sit at the intersection of AutoML and AI APIs allowing users to partially tailor the service to their needs. To ease discussion, here we focus on AI APIs that are offered 'as is' and do not allow for further customisation, though similar issues may well be relevant, particularly depending on the degree to which customisation is possible.

[3]Note there is much scepticism (which we share) about models for such tasks; we use age estimation as our exemplar as it is both a service offered by many AIaaS providers and represents an intuitive way to present the broader point.

Company	Asian	Black	Indian	White
Amazon	8.39	9.21	7.85	7.20
Baidu	5.34	7.89	6.90	7.54
Face++	6.72	8.74	8.25	8.62
Microsoft	5.26	5.88	5.76	5.19

Table 2: Age estimation performance across race groups as measured by Mean Absolute Error (MAE) metric.

Company	Asian		Black		Indian		White	
	u. 60	60+	u. 60	60+	u. 60	60+	u. 60	60+
Amazon	4.87	19.98	6.31	18.5	4.7	15.66	5.34	16.75
Baidu	3.69	10.78	5.44	15.72	4.58	14.75	5.46	14.94
Face++	6.07	8.82	8.52	9.44	8.2	8.43	9.08	6.99
Microsoft	3.44	11.22	4.33	11.17	4.03	9.15	4.19	11.21

Table 3: Age estimation performance across age and race groups as measured by Mean Absolute Error (MAE) metric.

Company	Asian					Black					Indian					White				
	<15	15-24	25-44	45-64	65+	<15	15-24	25-44	45-64	65+	<15	15-24	25-44	45-64	65+	<15	15-24	25-44	45-64	65+
Amazon	1.38	3.94	5.08	9.67	21.88	5.34	5.25	5.8	9.85	19.85	2.88	4.63	4.28	7.73	16.46	3.08	4.64	5.09	9.31	17.15
Baidu	1.24	1.83	4.22	8.34	11.11	3.29	4.9	4.95	10.07	16.23	2.46	4.8	4.13	7.94	15.12	2.5	6.06	4.63	9.46	15.06
Face++	3.19	6.04	7.27	8.17	8.76	13.06	5.48	6.77	8.66	9.81	11.37	6.11	7.02	8.45	8.32	13.59	6.86	8.12	7.58	6.92
Microsoft	1.1	2.61	4.18	6.09	12.15	2.09	3.73	4.71	7.39	11.8	1.31	5.21	4.35	5.33	9.64	1.92	4.65	4.39	6.31	11.42

Table 4: Age estimation performance across age and race groups as measured by Mean Absolute Error (MAE) metric.

and mitigating bias, fairness and other concerns of a model requires an understanding of *salient factors of the context in which it will be applied*.

From an AI API provider's perspective, who seeks to offer a generic 'one-size-fits-all' service, this therefore raises the general question of whether eliminating all types of bias and making the model representative of all groups is ever possible? Without knowing *how* their services will be used, *where, when* and by *whom*, despite best intentions, they might be unable to foresee all the various contexts in which their services will be used and thus likely fail to address the context-specific needs of a user use-case.

5.2.2 Provider perceptions. It is clear that the models a provider builds will be informed by or otherwise reflect (to varying degrees, depending on the circumstances) their view of the world. Given that AI APIs generally take a 'one-size-fits-all' approach, where such services are made widely and generally available, there will inevitably be mismatches between the perspectives and considerations of a provider and those users who seek to integrate these services to drive their applications.

For example, Terrance et al. [38] found that publicly available object-recognition services are less effective at recognising household items that are common in particular geographies and low-income communities. Suggested factors for this included differences in their appearance or the contexts in which they appear, and the use of 'English' as a base-language to describe particular items. This example represents a real-world case showing that globally-available AI services, reflecting the provider's own perspectives and assumptions, can under-perform when applied in particular user contexts (here around particular geographies and income-levels).

Further, we argue there are some AI API services that aim at areas that are too contextual, subjective, or ill-defined to form a general model for use by a range of actors, or indeed, even to be modelled at all. Using an extreme example to illustrate, we find that two prominent AIaaS providers (Baidu and Face++) now include a beauty estimation feature as part of their facial analysis offerings [13, 100], promoting it in cosmetics and matchmaking sectors, but the feature is available for general use. Beauty, however, is an inherently context-dependent and highly-subjective concept, which naturally varies between individuals, cultures, and regions. While beauty estimation services immediately raise red flags for a number of reasons, we mention them as they exist, are openly available commercially, and starkly illustrate how AIaaS providers can make certain assumptions or take certain positions regarding the world. That is, here, the providers of the service have decided on what 'beauty' looks like and how it should be scored – which will be reflected downstream in all their service user's applications. Interestingly, we see that Baidu's service outputs a single 'beauty-score' for any image, while Face++ gives two outputs to represent beauty scores "from both male's and female's perspectives" [100]; which illustrates not only how a provider's perspective impacts the services they offer, but also that these can differ from that of others, including other providers.

While it is generally the case that providers' views are encapsulated in their models (to varying degrees) for all services, this is especially problematic for concepts which are highly-subjective, and for which there is no generally agreeable 'ground-truth'. Naturally this is the case for beauty scoring, which is something that cannot be generally modelled, and in our view—not least given the potential ramifications of such a service—it would be wholly inappropriate to try.

The broader overarching point (inappropriate services aside) is that certain assumptions, values and ideals may hold in some contexts and situations but not in others. There will be many situations where a provider will not have considered the particulars of the social, economic, environment or other context of use; where a provider decides, judges or takes a position on particular concepts that are not or cannot be universally agreed upon; or where it is challenging, infeasible or inappropriate to reduce a range of distinct concepts, definitions and groups into a single model, let alone for models intended to operate generally (i.e. as a service) for use by others in a range of different contexts. Aiming to provide 'universal' functionality, which abstracts away or otherwise does not account for real-world differences will inevitably lead to problems.

5.3 Fully-managed AI Services

Finally, in the Fully-managed AI Services model, the control over the AI system is delegated almost entirely to a third party, who provides access to packaged solutions, typically entailing some particular workflow that incorporates one or more ML models. Some prominent examples of this service category include HireVue (algorithmic recruitment) [65], Luminance (AI for the legal profession) [94], and Lunit (AI for radiology) [95], all of which offer access to complete applications and systems that are driven by AI to varying extents. Similar to the AI APIs, the ML models underpinning fully managed services are often offered on an 'as-is' basis where no further user-led model retraining or larger customisation is possible, and are thus vulnerable to many of the challenges we have described in the previous section. However, the fully-managed service model contrasts with the AI API approach which typically concerns the more direct interactions between the user and access to the model itself, which is here further removed. Moreover, fully-managed services tend to deal with more numerous and/or complex set of functionalities—targeting niche industry-specific needs, in which contextual-specifics may be of even greater importance—potentially amplifying existing or incurring new kinds of risks and tensions. We elaborate these points below.

5.3.1 Representative of whom? We explored above how the 'one-size-fits-all' ideal underpinning AI APIs can lead to potential issues and tensions when these services are applied in the contexts their providers have not accounted for, or when the service aims to aggregate multiple inherently heterogeneous groups into one, resulting in a model that fails to accurately represent certain groups. Here, we argue that the complex, specialised subject matter of the fully-managed services and the generally opaque-manner in which they operate potentially amplify these risks in two key ways and make the question of their representativeness even more pertinent.

First, is the issue of the number of social categories and other factors that might be salient in deciding when determining a service's fairness, particularly as fully-managed services tend to embed a range of processes and workflows. One prominent area in which fully managed AI services are becoming increasingly common is healthcare – this includes services for radiology [95, 111, 124], dermatology [8], mental health [62, 145] among others. These services offer access to AI-enabled workflows and platforms, incorporating one or more pre-built ML models and potentially various other inputs, that allow users (customers) to predict certain health conditions [5, 95, 124], discover patterns and changes in individuals' health [62, 145] or receive medical recommendations and advice [62, 111]. Similar to the AI APIs, with the services already largely pre-built and allowing limited further user customisation, the providers of these services need to again make assumptions about who the target users of their services will be and what training data they should use to build their models. As we have shown, this is challenging even in the context of single AI APIs, such as the Facial Analysis services, where evaluations relate to a set of common physical attributes such as gender, skin type, and the intersections thereof. The challenges of representativeness are likely only be exacerbated in more complex domains, such as healthcare diagnostics, where there may be a virtually unlimited number of individual, group, and social factors that can be relevant, from lifestyle and eating habits, to weather and chemical exposure, and particularities regarding where one lives [24, 59, 119].

Secondly, and further complicating the matter is the issue of *testability*. Whereas AI API services are often openly accessible to anyone, which allows potential users as well as researchers and regulators to examine and validate these services to *some* extent (e.g. by validating their performance against different benchmarks and types of data [22, 133]), the access to fully managed AI services is by and large strictly limited to those directly engaging with the firm, or through provider-managed 'demos' which often come with limits or restrictions on functionality and use (and thus might effectively prevent interrogation). Moreover, whereas the API paradigm allows for direct interaction with the model itself, for fully-managed services often the workflows are often more complex, possibly involving several models and other processes (technical or otherwise) interacting with each other. This can make it harder to examine these services and the specific aspects of concern. In all, these characteristics limit the scope for potential external oversight as well as prospective user's ability to compare and validate the representativeness and suitability of these services for their particular application, exacerbating the risks of potential fairness conflicts and tensions.

5.3.2 Fair how? Data choices notwithstanding, it is also important to consider what does it mean for these systems to be *fair*? As an example consider algorithmic recruitment services (e.g. [65, 123]), which provide users (customers) with access to a suite of candidate assessments (questions, video interviews, games etc.) which are then algorithmically analysed by the provider and their outputs combined to score and rank candidates. These systems are concerned with allocating or withholding opportunities or resources; therefore, fairness is of key concern. But 'fair' can mean different things in different contexts to different people [109]. As we discussed in §2.3 scholars have proposed a number of competing ways to quantify fairness [46, 75, 93, 148], some of which are mathematically incompatible with each other [26, 51, 80]. In the typical machine learning development process, the creator of the system selects which definition of fairness to apply based on own understanding of the specific contextual needs of a given use case. However, this poses a unique challenge to the provider of an AI service because there may be multiple contexts and use cases for which different ways of measuring fairness may be applicable (see §5.1.2, §5.2.1). It also poses a challenge to the service user, who may not have all the information to assess for fairness. The user may have access to general information about candidates scores and the overall ranking, for instance, but the full details of the selection process, fairness metrics and mitigation measures used, tend to be managed by the provider and hidden from direct user oversight and control.

This can lead to potential tensions, where the definition of fairness chosen by the service provider is incompatible with the notion of fairness held by the customer, or the legal requirements of the market they operate in. For instance, in the now well-known case of COMPAS recidivism risk assessment software, ProPublica's investigative journalists argued that the system is biased against black defendants based on its failing of the "False Positive Rate Parity" definition of fairness [10]. In response, the creator of the system

(Northpointe) argued that its system was in fact fair according to "Positive Predictive Value Parity" definition of fairness, and it is the measure of fairness that ProPublica selected that was flawed [10]. Similar issues can arise in other fully managed AI services, where user needs and expectations may not align with, or indeed, be incompatible with the assumptions and approaches taken by the provider.

As it stands, without defined sets of standards or laws on which fairness metrics need to be met and with the customers of AI services having no say in the definition of fairness to be followed, there is a risk that service providers might thus opt to adopt fairness definitions according to their own concerns, priorities, and interests. For instance, in their evaluation of commercial algorithmic employment assessment services [125], Raghavan et al. found that all evaluated providers who made concrete claims about fairness and debiasing practices of their systems did so with reference to the U.S. Equal Employment Opportunity Commission (EEOC) Uniform Guidelines *4/5ths rule* [19]. This is understandable considering that majority of these providers are located in the U.S. and might thus be legally obliged to adhere to this rule. However, the customer base of these services is global and subject to wide range of different jurisdictions which may or may not follow similar guidelines. For instance, in the EU and UK, where no equivalent of the 4/5ths rule exists in either statute or case law and where concepts of direct and indirect discrimination raise different considerations around algorithmic fairness [2, 132], the provider's choice of 4/5ths rule as a measure of disproving bias is no longer so obvious, as other measures might be more appropriate. Given the global nature of services, and the multitudes of separate, potentially conflicting national guidelines and regulations to follow, providers will most likely consider the legal requirements of their own or key customers jurisdictions, but are unlikely to consider the full range of similar, overlapping and conflicting notions of bias, fairness, and discrimination in philosophical, sociological, legal and cultural context of every potential customer [136]. Again, by taking away the control of such context-dependent aspects from the customer, this fully managed AI service model risks imposing providers' world views and values onto their clients, which could lead to potential unintended consequences and tensions.

6 DISCUSSION

As AIaaS grows in prominence, so too does the risk of the real-world harm of these services. Indeed, as we have argued, there exists an inherent tension between the purportedly generic, 'one-size-fits-all' ideal underpinning AIaaS and the nuanced, context-dependent nature of fairness. In aiming to offer AI services generically to potentially millions of customers, AIaaS providers need to make certain assumptions or choices about where and how their services will be used and by whom. At the same time, the world is vast and full of diverse groups that practise a range of traditions, customs and activities and share a variety of views and beliefs. As we have shown in this work, inevitably there will be situations in which the 'generic' assumptions underpinning AI services will fail to address or come into conflict with the specific, context-dependent needs of a specific user's use case (which the AIaaS providers are unlikely to have considered for all their potential customers) and in turn may render these services and their end-users' systems biased.

We thus argue for caution and consideration in the development and use of AI services in practice. The AIaaS model has the potential to make AI more widely accessible, making it easier, faster and cheaper to deliver state-of-the-art AI and its benefits to a much wider range of applications and groups than might otherwise be possible. However, without proper care, these seeming benefits of 'turn-key' AI availability and virtually limitless scalability can quickly risk potentially exacerbating the many bias problems and the broader ethical concerns around AI that various research communities have repeatedly cautioned against.

Our focus so far has been on the risks and shortcomings of the AIaaS paradigm to draw attention to the area. We next identify potential ways forward and explore some considerations representing opportunities for discussion and future research in this space.

6.1 Provider Considerations

6.1.1 Appropriateness of the AI service. As we have detailed and shown, there are many situations in which the 'one-size-fits-all' ideal of AIaaS can come into conflict with the contextual needs of a user's application, leading to potential risks and harms. Therefore, we join the growing number of calls urging researchers and developers to shift to a mindset of careful planning and evaluation in order to determine whether an attempt to build an AI model—or in this case an AI service—is appropriate in the first place [15, 106, 128, 136]. While this concern is relevant for any ML model development, it is particularly pertinent in the AIaaS context where potential harms might achieve a far larger scale [29, 72]. Before building a new AI service, it is important to consider to what extent (if at all) is it possible for the service to accurately model the social and technical requirements of the various contexts in which it will be deployed. *Who* might be the target audiences and *how* might they use the service? *When* and *where* might the service possibly be used (or misused [71])? Is it feasible or even possible for the model to accurately capture the social environment and its various needs, or would the modelling required be so complex as to be computationally intractable? Or indeed, is the topic too subjective or ill-defined to be one that is suitable or appropriate for general modelling?

As an illustrative example, an AI service may be designed for identity verification for users signing up for new financial products. The service providers should consider the possibility that this could be adapted for other purposes—such as age prediction for adult content, or policing and surveillance—and ask themselves whether this is appropriate and acceptable. Possible mitigative actions on the vendor's part may include 1) disclosure on the original purpose of the model, 2) documentation on the data and model, such as its representativeness in age, gender, and nationality, and 3) contractual and license restrictions for certain usage, e.g. cannot be used for illegal activities, for police surveillance, or even for some undesirable (but legal) purposes. While this may not prevent malicious actors from misusing the tool, the exercise of considering the potential alternative uses of the AI service could bring to light the contextual limitations to enable more effective disclosure and guidance to the users.

Furthermore, we propose that stakeholders involved in the development of AI services think through the potentially negative and harmful consequences that might arise from their services being applied in the context they have unaccounted for. As we discussed in §2.2 the EU's proposed AI Act provides for risk management processes for *certain* "high risk" AI systems [30]; we propose that similar processes should be required for *all AI services* due to the inherent risks of mismatch between systems' development by providers and deployment by users. If at any point, such a list of assumptions, caveats and limitations becomes overwhelming, then it could be a reasonable indication that the problem the given service aims to address is too contextual to be abstracted into a 'generic' AI service and consequently stop the development of such service.

6.1.2 Transparency. As a part of the internal and external accountability processes, transparency can assist in facilitating algorithmic fairness. Apart from the role played by industry-wide regulations and standards, transparency mechanisms [17, 47, 131] should be embraced and implemented by developers and organisations committed to the development of AI services. Transparency which supports broader accountability processes on training sets, models and internal processes behind these services, as well as in the wider-sense of acknowledgement of known-limitations, past failures, and lessons learnt [27, 131] can help move AIaaS from the opaque, unknown and potentially harmful 'black-boxes' of today, to a more understandable, accepting of their own limitations, building blocks for affordable and widely accessible AI systems of the future.

We note that just as it is recognised that transparency won't necessary solve algorithmic accountability issues [9, 85, 127], transparency will not by itself address all AIaaS concerns. However, by embracing transparency at every stage of the AIaaS development process, including being upfront about the design contexts considered and any known limitations of the service while marketing it to potential users, some of the risks we have cautioned against in this paper could be reduced. There appears a clear role for adopting and extending various documentation approaches, be they relating to data and models (such as datasheets [53], model cards [106] and factsheets [12]), their interactions and interconnections (e.g. decision provenance [139]), and broader, holistic socio-technical system reviews (e.g. reviewability [27] and traceability [86]).

Similarly, from providing varying levels of control and service-details depending on the user's proficiency with ML (particularly within the AutoML context) [157], to partnering with external auditors [39, 84, 126] to help uncover potential gaps and limitations of the service, there are many opportunities for (research on) increasing transparency and improving collaboration between developers, customers, and stakeholders of AIaaS models and services. Importantly, these mechanisms should connect to internal and external institutional and governance mechanisms for account-giving and redress to ensure that issues are properly identified and addressed (whether to compliance and assurance teams within providers, to regulators and other external oversight bodies, or to customers of AI services who wish to ensure that required standards are being met in the services they are using).

6.1.3 Gatekeeping. It is worth exploring the options for providers in taking a more active role in setting out the intended usage of their services and mitigating the risks stemming from the repurposing of their services in untested or untoward contexts. For example, they might wish to limit their services to users from particular regions, jurisdictions or industries, which they believe best fit the capabilities and known limitations of their service, thus avoiding some of the risks in dealing with unknown contexts. Further, moving beyond just the issue of fairness, AIaaS providers might wish to monitor and restrict the usage of their services in order to mitigate the reputational risks and issues that could arise from the misuse of their services in driving controversial, inappropriate, or even illegal applications – as Javadi et al. discuss [71, 72].

Conversely, one could question whether it is sensible for AIaaS providers to decide on what constitutes an appropriate, ethical or untoward use case, or who should or should not have access to certain set of services. Though here we do not explore the various business, legal and other perspectives that go behind this concept of gatekeeping, we highlight this as an area requiring further consideration.

6.2 Usage Considerations

6.2.1 Responsible AIaaS procurement and usage. In §6.1.1 we argued that before building a new AI service, AIaaS providers should first consider the potential risks and limitations to determine *whether* and *if* such developments should be undertaken at all. Similarly, we argue that the users of AIaaS services should also carefully consider the appropriateness and potential risks of procuring AIaaS solutions to drive their applications. Notably, it is the users of AIaaS that have the advantage of being placed to know the precise context in which the AI service will be used, something the AIaaS providers are currently only generally able to try to anticipate. It is therefore worth exploring the particular ways in which AIaaS users could leverage this knowledge to their advantage, so to support and effectively empower them in assessing the suitability of an AI service in the context of their applications.

Central to this will be the service's transparency mechanisms which we have discussed in the previous section, whereby information is made available to service users. This could include transparency on a technical level such as the training data used, fairness metrics and measures applied, performance achieved or methods adopted, and the interactions between models; but also transparency in a more general sense such as details of their development practices, intended usage and known limitations [125]. However, for this to work, it is crucial that the transparency provided is meaningful and effective for users, and thus raises opportunities for the HCI community to play an important role in realising these.

Additionally, for some AI services—particularly those of an AutoML or AI API type—there might be scope for potential customers to take on a more proactive role and examine the suitability of a given service through their own algorithmic investigations. As we have shown in this paper, it may be difficult (if not impossible) for AIaaS providers to exhaustively test and validate their services as free from all kinds of bias in all kinds of contexts [127, 136]. However, by conducting their own experimentation—not unlike the experiments we presented in §5.1 and §5.2—as part of their provider/service vetting process, the users of AI services could themselves be well-placed to validate service performance *in context of their*

specific use case. Indeed, a recent line of HCI research has explored the idea of "everyday algorithm audits" whereby users detect, understand, and interrogate problematic machine behaviours via their day-to-day interactions with algorithmic systems [39, 43, 138]. We argue that there will be benefits in including AIaaS in those discussions; for example, by exploring designs and methods that better support users in conducting more effective service audits for bias, and means for meaningfully presenting and communicating such findings back to service providers. In practice, this will require understanding service users' needs as it regards fairness testing, providing effective interfaces and usable tools for such, and so on.

6.2.2 Understanding user needs and challenges around AIaaS usage. Despite growth in the development and dissemination of AIaaS services, there has been little research investigating how AIaaS customers *actually* use these services in practice, nor on practitioners' experiences and desires around AIaaS and the gaps between the capabilities of existing AI services and the needs of their users. Similarly, little is known about who the typical AIaaS user is, what level of ML-proficiency do they possess in practice, or in which markets they operate in. This limits the accessibility and usability of the AI service and presents a risk that these services and any supporting tools might be designed in a way that fails to align with user needs.

These issues have been shown in more general ML-contexts; for example, studies exploring user (ML practitioner) attitudes and expectations regarding fairness toolkits showed a misalignment of tools with user expectations and needs, and generally being difficult to use (among other concerns), which not only makes such tooling less-effective, but can potentially lead to increased bias risks [40, 91]. Similarly, in one of the few studies exploring practitioner experiences using AutoML tools and platforms, Xin et al. demonstrated a dissonance between those building AutoML tooling seeking to achieve "full automation" and that of actual users for whom "a complete automation is neither a requirement nor a desired outcome" [157]. Instead, they found that users expressed interest in a more "human-in-the-loop" AutoML—that aims to empower users in undertaking ML development, rather than replace them completely.

As such, we argue that the AIaaS space would benefit greatly from the attention of the HCI community, by studying how practitioners actually attempt to use these services in the context of real-world tasks. This could also include exploring their prior knowledge and understanding of ML processes, awareness of legal obligations, attitudes towards fairness, amongst other concerns. Such understandings might result in highlighting new opportunities for tooling, or devising a common vocabulary and effective modes of presentation for the providers and users of AIaaS to discuss and compare individual services, communicate issues, and assist users in identifying provider services that best align with the views and needs of their specific application contexts. In general terms, there seems much scope for users to be better engaged as part of the design of AI services, which ultimately can help better support the proper use of AIaaS and make the benefits of AI more widely-accessible, while reducing the potential risks. On the other hand, poorly implemented or poorly explained AI services risk that a customer would select a suboptimal or inappropriate tool for their use case, or simply decide against adopting AIaaS at all. In all, there are clear opportunities for research in this space.

6.3 Policy Directions

6.3.1 Roles and responsibilities. As we have discussed, there are increasing moves towards regulating AI systems, including the EU's proposed AI Act which will apply to "providers" of AI services. As we note in §2.2, however, AIaaS challenges established understandings of the roles and responsibilities of providers and users; in some cases, the users (customers) of those services may themselves become considered as the "provider" under the Regulation for the underlying AI system in relation to their use of the service. As such, the compliance obligations (which include those relating to bias and fairness) applying to providers will in those cases also apply to these "user-providers" (Article 28, EU AI Act). Yet such user-providers are downstream of the problems that the AI Regulation [31] seeks to address. As this paper makes clear, in a service context such-user providers will therefore lack anything like the information needed to meet their obligations; may not be in a position to understand or deal with such concerns, despite provider claims of "*no expertise required*"; nor will they have the technical influence over such systems to be able to do so.

Moreover, as we argue in this paper, general purpose models underpinning AIaaS can often exhibit structural issues that cannot be simply tweaked out of existence downstream by user-providers. Any biases and other undesirable characteristics ingrained in these services (by the AIaaS provider) will thus inevitably propagate downstream and there may be little downstream user-providers can do to address this; which is particularly problematic where—as this work makes clear—even systems that seem to be generically 'fair' under some (generally broad, sweeping) measure, may in fact exhibit bias issues when deployed in a specific context. As it stands, the downstream user-providers are given no help in scrutinising or holding to account upstream providers and instead are tasked with impossible compliance tasks. We therefore argue that the role of AIaaS providers and customers is an important issue warranting consideration in regulation targeting issues of AI fairness.

6.3.2 AIaaS regulation and guidance. The emergence of the AIaaS model and its growing popularity calls for its explicit inclusion in ongoing debates on AI regulation, potentially necessitating the development of AIaaS-specific regulation, standards, and guidance. Indeed, given current regulatory directions [31, 49, 74], there appears both scope and appetite for oversight bodies to set directions regarding various issues relating to AIaaS. The consolidation of AI services around a few providers and the potential future widespread use of those services—although bringing challenges, as we have argued—may also mean that actually targeting specific regulation, oversight, and enforcement mechanisms at certain AI-related problems becomes easier than if many companies were developing their own AI systems in house.

For instance, regulations could set limits on acceptable types and uses of AI services and around monitoring for misuse of AI services for illegal or potentially harmful purposes, a recognised concern [71, 72]. Such regulations could specify the role for various relevant actors—such as regulators, industry, and civil society—in complying with and overseeing these limits. In doing so, regulation

may require explicit bans on types or uses of AI services which are deemed to pose a high risk to the rights, freedoms, and interests of people potentially affected by them, or more lenient sets of requirements, standards, and best practices for other, less risky systems. These could relate to, for example, the production and suitability of training data, to training and testing procedures and other areas of systems' development, as well as other aspects of a service offering with a view to identifying and addressing biases, in a manner that properly accounts for the position of the user as a customer of such services.

6.3.3 Facilitating regulatory oversight. As we have argued, transparency and accountability regimes are also needed to provide the information, mechanisms, and processes that allow the design and functioning of AI services to be better understood and to facilitate effective challenge and oversight by regulators, users (i.e. customers), and others. Mechanisms that enable regulators to obtain further information where needed would also help, as would mechanisms enabling users and regulators to hold providers to account and take corrective action where needed. While there is a growing interest in auditing algorithmic systems [125, 126, 128, 138, 154], this field generally lacks defined standards and records could be incomplete or not relating to information needed for accountability purposes [15, 127]. There may therefore be a role for regulators in defining standards and ensuring consistency in auditing practices so as to assist with facilitating oversight, particularly those cognisant of the 'supply chain' that AIaaS inherently entails, and the nature of the relationships of the actors within that. Indeed, such activities also present opportunities for the HCI community, leveraging and extending work on how to provide means and tooling for meaningful transparency to support accountability [11, 27, 47, 63, 110].

While we concretely highlight issues and provide suggestions to indicate some potential avenues forward, the specifics regarding external oversight and scrutiny over the use, reliance, monitoring and actions of AI services require further consideration. Future research of an academic and policy nature may explore these issues in greater depth so as to develop more proposals for intervention in this space.

7 CONCLUSION

In this paper we explored and elaborated a range of potential fairness concerns posed by AIaaS. Specifically, we introduced, reviewed and systematised the AI services landscape, proposing a general taxonomy of service type characteristics from a service user perspective. Further, we have examined each of the different types of AIaaS services using a combination of experimental evaluation and illustrative examples, outlining the bias and fairness concerns and practical challenges they bring to the 'fair ML' discussion. We have highlighted the broader implications that the emergence of AIaaS has for issues of fairness and AI governance, suggesting potential research opportunities and ways forward.

The adoption of AIaaS will likely only continue to grow. As such, it is critical to address the open challenges, limitations and concerns this service paradigm presents. By providing this initial exploration of AIaaS fairness considerations, we hope to raise awareness of the variety of challenges in this space, and foster a discussion on what can be done to mitigate these risks.

ACKNOWLEDGMENTS

We acknowledge the financial support of the UKRI Engineering and Physical Sciences Research Council (EPSRC) (EP/P024394/1 and EP/R033501/1), Aviva, and Microsoft through the Microsoft Cloud Computing Research Centre.

REFERENCES

[1] 4Paradigm. 2022. 4Paradigm Sage HyperCycle ML. https://www.4paradigm.com/product/hypercycleml.
[2] Jermias Adams-Prassl, Reuben Binns, and Aislinn Kelly-Lyth. 2022. Directly Discriminatory Algorithms. *Modern Law Review* (2022). https://doi.org/10.1111/1468-2230.12759
[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.
[4] Nur Ahmed and Muntasir Wahed. 2020. The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research. *arXiv preprint arXiv:2010.15581* (2020).
[5] Aidence. 2022. AI-powered clinical applications for the oncology pathway. https://www.aidence.com/.
[6] Amazon. 2021. Amazon SageMaker Autopilot. https://aws.amazon.com/sagemaker/autopilot/.
[7] Amazon. 2021. Artificial Intelligence Services. https://aws.amazon.com/machine-learning/ai-services/.
[8] Skin Analytics. 2022. Providing AI supported dermatology solutions in partnership with the NHS. https://skin-analytics.com.
[9] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society* 20, 3 (2018), 973–989. https://doi.org/10.1177/1461444816676645
[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2019. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. 2016. *URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* (2019).
[11] Ariful Islam Anik and Andrea Bunt. 2021. Data-centric explanations: explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[12] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.
[13] Baidu. 2021. Baidu AI Open Platform. https://ai.baidu.com/tech/face/detect.
[14] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.
[15] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.
[16] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI conference on Human Factors in Computing Systems*. 1–14.
[17] Clare Birchall. 2014. Radical transparency? *Cultural Studies? Critical Methodologies* 14, 1 (2014), 77–88.
[18] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).
[19] Philip Bobko and Philip L Roth. 2004. The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice. In *Research in personnel and human resources management*. Emerald Group Publishing Limited.
[20] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).
[21] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[22] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, Vol. 81. PMLR, 77–91.

[23] Corinne Cath. 2018. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. , 20180080 pages.

[24] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. 2019. Artificial Intelligence, bias and clinical safety. *BMJ Quality & Safety* 28, 3 (2019), 231–237.

[25] Joseph Chin, Aifaz Gowani, Gabriel James, and Matthew Peng. 2020. The death of data scientists - will automl replace them? https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html

[26] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[27] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 598–609.

[28] Jennifer Cobbe, Chris Norval, and Jatinder Singh. 2020. What lies beneath: transparency in online service supply chains. *Journal of Cyber Policy* 5, 1 (2020), 65–93. https://doi.org/10.1080/23738871.2020.1745860

[29] Jennifer Cobbe and Jatinder Singh. 2021. Artificial intelligence as a Service: Legal responsibilities, liabilities, and policy challenges. *Computer Law & Security Review* 42 (2021), 105573.

[30] European Commission. 2021. Proposal for regulation of the European parliament and of the council - Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[31] EU Commission et al. 2021. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *COM (2021)* 206 (2021).

[32] McKinsey & Company. 2020. The state of AI in 2020. https://www.mckinsey.com/Business-Functions/McKinsey-Analytics/Our-Insights/Global-survey-The-state-of-AI-in-2020.

[33] Kate Crawford. 2016. Artificial intelligence's white guy problem. *The New York Times* 25, 06 (2016).

[34] André F Cruz, Pedro Saleiro, Catarina Belém, Carlos Soares, and Pedro Bizarro. 2020. A Bandit-Based Algorithm for Fairness-Aware Hyperparameter Optimization. *arXiv preprint arXiv:2010.03665* (2020).

[35] Marija Cubric. 2020. Drivers, barriers and social considerations for AI adoption in business and management: A tertiary study. *Technology in Society* 62 (2020), 101257.

[36] Allan Dafoe. 2018. AI governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK* (2018).

[37] Thomas Davenport and Ravi Kalakota. 2019. The potential for artificial intelligence in healthcare. *Future healthcare journal* 6, 2 (2019), 94.

[38] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 52–59.

[39] Wesley Hanwen Deng, Bill Boyuan Guo, Alicia Devos, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2022. Understanding Practices, Challenges, and Opportunities for User-Driven Algorithm Auditing in Industry Practice. *arXiv preprint arXiv:2210.03709* (2022).

[40] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. (2022), 473–484. https://doi.org/10.1145/3531146.3533113

[41] Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. 2019. Image counterfactual sensitivity analysis for detecting unintended bias. *arXiv preprint arXiv:1906.06439* (2019).

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[43] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[44] Yogesh K Dwivedi, Laurie Hughes, Elvira Ismagilova, Gert Aarts, Crispin Coombs, Tom Crick, Yanqing Duan, Rohita Dwivedi, John Edwards, Aled Eirug, et al. 2019. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management* (2019), 101994.

[45] Yogesh K Dwivedi, Laurie Hughes, Elvira Ismagilova, Gert Aarts, Crispin Coombs, Tom Crick, Yanqing Duan, Rohita Dwivedi, John Edwards, Aled Eirug, et al. 2021. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management* 57 (2021), 101994.

[46] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the Third Innovations in Theoretical Computer Science Conference*. 214–226.

[47] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[48] Hugging Face. 2022. The AI community building the future. https://huggingface.co.

[49] Federal Trade Commission, USA. 2021. Aiming for truth, fairness, and equity in your company's use of AI | Federal Trade Commission. https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai.

[50] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[51] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[52] Megan Garcia. 2016. Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal* 33, 4 (2016), 111–117.

[53] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[54] Markos Georgopoulos, James Oldfield, Mihalis A Nicolaou, Yannis Panagakis, and Maja Pantic. 2021. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision* 129, 7 (2021), 2288–2307.

[55] Google. 2021. Cloud AI Building Blocks. https://cloud.google.com/products/ai/building-blocks.

[56] Google. 2022. Cloud AutoML Custom Machine Learning Models | Google Cloud. https://cloud.google.com/automl.

[57] Ben Green and Lily Hu. 2018. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning (ICML)*.

[58] Patrick Grother, Mei Ngan, and Kayee Hanaoka. 2019. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. National Institute of Standards and Technology.

[59] Stanford HAI. 2022. The Geographic Bias in Medical AI Tools. https://hai.stanford.edu/news/geographic-bias-medical-ai-tools.

[60] Mark Haranas. 2022. AWS, Microsoft, Google Top Cloud Ai Developer Market: Gartner. https://www.crn.com/slide-shows/cloud/aws-microsoft-google-top-cloud-ai-developer-market-gartner

[61] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).

[62] Headspace. 2022. Headspace Health. https://www.headspace.com/health.

[63] Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–29.

[64] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 771–787.

[65] HireVue. 2022. End-to-End Hiring Experience Platform: Video Interviewing, Conversational AI & More | HireVue. https://www.hirevue.com/.

[66] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[67] Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and Engineering Ethics* 24, 5 (2018), 1521–1536.

[68] IBM. 2021. IBM Watson products and solutions. https://www.ibm.com/uk-en/watson/products-services.

[69] IBM. 2022. Watson OpenScale on Cloud Pak for Data. https://www.ibm.com/docs/en/cloud-paks/cp-data/3.5.0?topic=services-watson-openscale.

[70] Infermedica. 2022. Call Center Triage. https://infermedica.com/product/call-center-triage.

[71] Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2020. Monitoring Misuse for Accountable 'Artificial Intelligence as a Service'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 300–306.

[72] Seyyed Ahmad Javadi, Chris Norval, Richard Cloete, and Jatinder Singh. 2021. Monitoring AI Services for Misuse. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 597–607.

[73] Haifeng Jin, Qingquan Song, and Xia Hu. 2019. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD international Conference on Knowledge Discovery & Data Mining*. 1946–1956.

[74] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[75] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *arXiv preprint arXiv:1605.07139* (2016).

[76] Dimitra Kamarinou, Christopher Millard, Jatinder Singh, and R Leenes. 2017. Machine learning with personal data. In *Data protection and privacy: the age of intelligent machines*. Hart Publishing.

[77] Zaid Khan and Yun Fu. 2021. One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 587–597.

[78] Udayan Khurana, Horst Samulowitz, and Deepak Turaga. 2018. Feature engineering for predictive modeling using reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[79] Ross D. King, Cao Feng, and Alistair Sutherland. 1995. Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence an International Journal* 9, 3 (1995), 289–333.

[80] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[81] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.

[82] Ron Kohavi. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.. In *Kdd*, Vol. 96. 202–207.

[83] Lars Kotthoff, Chris Thornton, Holger H Hoos, Frank Hutter, and Kevin Leyton-Brown. 2019. Auto-WEKA: Automatic model selection and hyperparameter optimization in WEKA. In *Automated Machine Learning*. Springer, Cham, 81–95.

[84] PM Krafft, Meg Young, Michael Katell, Jennifer E Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, et al. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 772–781.

[85] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. Accountable algorithms. *University of Pennsylvania Law Review* 165, 3 (Feb. 2017), 633–705.

[86] Joshua A. Kroll. 2021. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 758–771. https://doi.org/10.1145/3442188.3445937

[87] Hoang Thanh Lam, Johann-Michael Thiebaut, Mathieu Sinn, Bei Chen, Tiep Mai, and Oznur Alkan. 2017. One button machine for automating feature engineering in relational databases. *arXiv preprint arXiv:1706.00327* (2017).

[88] Doris Jung Lin Lee, Stephen Macke, Doris Xin, Angela Lee, Silu Huang, and Aditya G Parameswaran. 2019. A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. *IEEE Data Eng. Bull.* 42, 2 (2019), 59–70.

[89] Michelle Seng Ah Lee and Luciano Floridi. 2021. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines* 31 (2021), 165–191.

[90] Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. 2021. Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. *AI and Ethics* 1, 4 (2021), 529–544.

[91] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[92] Daphne Leprince-Ringuet. 2021. Low-code and no-code development is changing how software is built - and who builds it. https://www.zdnet.com/article/low-code-and-no-code-development-is-changing-how-software-is-built-and-who-builds-it/

[93] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875* (2017).

[94] Luminance. 2022. The Artificial Intelligence platform for the legal profession. https://www.luminance.com.

[95] Lunit. 2022. AI will be the new standard of care. By Lunit. https://www.lunit.io/en.

[96] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[97] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[98] Gustavo Malkomes, Chip Schaff, and Roman Garnett. 2016. Bayesian optimization for automated model selection. In *Workshop on Automatic Machine Learning*. PMLR, 41–47.

[99] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*. PMLR, 6755–6764.

[100] MEGVII. 2021. Megvii Face++ Artificial Intelligence Open Platform. https://www.faceplusplus.com.cn/beauty/.

[101] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[102] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[103] Microsoft. 2021. Cognitive Services. https://azure.microsoft.com/en-gb/services/cognitive-services/.

[104] Microsoft. 2022. Automated Machine Learning | Microsoft Azure. https://azure.microsoft.com/en-gb/services/machine-learning/automatedml/.

[105] Microsoft. 2022. Our approach to responsible AI at Microsoft. https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1:primaryr5.

[106] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[107] ModelZoo. 2022. Discover open source deep learning code and pretrained models. https://modelzoo.co.

[108] Jakob Mökander, Prathm Juneja, David S Watson, and Luciano Floridi. 2022. The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other? *Minds and Machines* (2022), 1–8.

[109] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–36.

[110] Chris Norval, Kristin Cornelius, Jennifer Cobbe, and Jatinder Singh. 2022. Disclosure by Design: Designing information disclosures to support meaningful transparency and accountability. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 679–690.

[111] Nuance. 2022. AI Marketplace for Diagnostic Imaging - AI for Radiology. https://www.nuance.com/healthcare/diagnostics-solutions/ai-marketplace.html.

[112] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[113] OECD. 2019. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[114] United Nations. Statistical Office. 1982. *Provisional guidelines on standard international age classifications*. New York: United Nations.

[115] Randal S Olson and Jason H Moore. 2016. TPOT: A tree-based pipeline optimization tool for automating machine learning. In *Workshop on Automatic Machine Learning*. PMLR, 66–74.

[116] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[117] OpenAI. 2022. ChatGPT. https://openai.com/blog/chatgpt/.

[118] OpenAI. 2022. DALL-E 2. https://openai.com/dall-e-2/.

[119] Seong Ho Park and Kyunghwa Han. 2018. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 286, 3 (2018), 800–809.

[120] Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. 2021. Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 854–863.

[121] Florian Pfisterer, Stefan Coors, Janek Thomas, and Bernd Bischl. 2019. Multi-objective automatic machine learning with autoxgboostmc. *arXiv preprint arXiv:1908.10796* (2019).

[122] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *arXiv preprint arXiv:1709.02012* (2017).

[123] Pymetrics. 2021. Talent Matching Platform. https://www.pymetrics.ai/.

[124] qure.ai. 2022. AI to enable accessible, affordable & timely care across the globe. https://qure.ai/.

[125] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.

[126] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

[127] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 145–151.

[128] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.

[129] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[130] Grand View Research. 2021. Artificial Intelligence as a Service: Market Research Report. https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-as-a-service-market-report.

[131] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 315–328.

[132] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 2020. What does it mean to'solve'the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 conference on Fairness, Accountability, and Transparency*. 458–468.

[133] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[134] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–35.

[135] Daniel Schiff, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. What's Next for AI Ethics, Policy, and Governance? A Global Overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 153–158.

[136] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[137] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).

[138] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[139] Jatinder Singh, Jennifer Cobbe, and Chris Norval. 2019. Decision Provenance: Harnessing Data Flow for Accountable Systems. *IEEE Access* 7 (2019), 6562–6574. https://doi.org/10.1109/ACCESS.2018.2887201

[140] Jatinder Singh, Ian Walden, Jon Crowcroft, and Jean Bacon. 2016. Responsibility & machine learning: Part of a process. *Available at SSRN 2860048* (2016).

[141] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2459–2468.

[142] Stability.AI. 2022. Stable Diffusion. https://stability.ai/blog/stable-diffusion-v2-release.

[143] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).

[144] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* (2019).

[145] Feel Therapeutics. 2022. Decoding Mental Health. https://www.feeltherapeutics.com.

[146] Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. 2019. Considerations for AI fairness for people with disabilities. *AI Matters* 5, 3 (2019), 40–63.

[147] US House of Representatives. 2022. H.R.6580 - Algorithmic Accountability Act of 2022. https://www.congress.gov/bill/117th-congress/house-bill/6580.

[148] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FAIRWARE)*. IEEE, 1–7.

[149] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *arXiv preprint arXiv:2005.05906* (2020).

[150] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[151] Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J Smith, Kalyan Veeramachaneni, and Huamin Qu. 2019. Atmseer: Increasing transparency and controllability in automated machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[152] Daniel Karl I Weidele, Justin D Weisz, Erick Oduor, Michael Muller, Josh Andres, Alexander Gray, and Dakuo Wang. 2020. AutoAIViz: opening the blackbox of automated artificial intelligence with conditional parallel coordinates. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 308–312.

[153] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 1–18.

[154] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 666–677.

[155] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[156] Qingyun Wu and Chi Wang. 2021. Fair AutoML. *arXiv preprint arXiv:2111.06495* (2021).

[157] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. 2021. Whither automl? understanding the role of automation in machine learning workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[158] Renee Yao. 2022. Lunit, maker of FDA-cleared AI for cancer analysis, goes public. https://blogs.nvidia.com/blog/2022/07/21/lunit-healthcare-ai-ipo/

[159] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*. 1171–1180.

[160] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[161] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al. 2021. The AI Index 2021 Annual Report. *arXiv preprint arXiv:2103.06312* (2021).

[162] Song Yang Zhang Zhifei and Qi Hairong. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[163] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).