AI and Ethics
https://doi.org/10.1007/s43681-023-00316-2

ORIGINAL RESEARCH

Technical challenges and perception: does AI have a PR issue?

Marie Oldfield[1]

Received: 22 September 2022 / Accepted: 24 June 2023

Abstract
Increasingly, models have been highlighted that not only disadvantage society but those whom the model was originally designed to benefit. An increasing number of legal challenges around the world illustrates this. A surge of recent work has focussed on the technical, legal or regulatory challenges but not necessarily the real-world day to day challenges for practitioners such as data collection or fairness by design. Since the publication of the Holstein et al.'s study in 2019, additional legislation, regulation and multiple bodies have been created to address practitioner challenge. This study asks what, if anything, has improved for practitioners between 2019 and 2022. Study 1 conducts an investigation into real-world needs within industry and asks whether practitioners are now able to mitigate challenges in a more robust manner. A further pilot study on the perception of AI examines whether perception of AI impacts practitioner work. The results show increasing and continuing interdisciplinary issues. Where increased regulation and legislation might have seemed reasonable, the result for practitioners is indecision and overwhelm. Based on these findings, we highlight directions for future research in this area. The most problematic area being human factors.

Keywords Machine learning · AI · User perception · Model development · Impact on society · Organisational behaviour

1 Introduction

Technology is inextricably linked to society and can have detrimental impacts on people when implemented in a non-ethical fashion. Many decisions, normally made by a human, are now becoming automated. Whilst some processes can be reduced in this way, some cannot. Decisions made about who can have a credit card, for example, were thought to be straightforward enough to automate until an algorithm, designed and implemented to make these decisions [8, 77, 103], was seen to be discriminatory. Decisions, whether it be what school a child goes to or whether one is eligible for a credit card or loan [93], ultimately shape society. From providing the evidence to support major investment decisions [78] to individual services for citizens [16, 19, 37, 56, 73, 84], models can have major implications for society. One example is the Homes Office Visa Algorithm that was implemented and then withdrawn after causing a huge amount of distress to those waiting or applying for visas. This incurred a substantial loss of taxpayers' money [7]. It is critical that models are fit for purpose. We must balance key attributes such as societal impact against supporting innovation, so that society can benefit from science without being exploited or discriminated against [6, 49, 71, 72]. Limiting the potential harms associated with poorly designed modelling is challenging but not impossible [25, 63, 103].

In this paper, we present two studies. In the first study, we investigate the challenges faced by commercial developer teams in Machine Learning (ML) and Artificial Intelligence (AI)[1] products and service teams, whose products have global impact. This is achieved through a re-run of the survey undertaken by Holstein et al. [44] in 2019 in USA.

[1] AI in this paper is used in the way developers and users understand the term. In academia, we might describe much of what is termed as 'AI' as machine learning or statistical modelling.

✉ Marie Oldfield
M.oldfield@lse.ac.uk

[1] LSE: The London School of Economics and Political Science, London, UK

Published online: 24 July 2023

We aim to determine what, if any, progress has been made since this study was published. Using JISC online surveys, social media, mailing lists and specialist bodies, we investigated the challenges faced by commercial ML/AI teams as well any needs for additional support or development of best practise. We identified a range of real-world needs. We also found that many of the challenges that were identified by Holstein et al. [44] still remain. This study also highlighted some substantially different findings from Holstein at al. [44] in that additional further real-world challenges were highlighted outside of the traditional model development pipeline, such as customer perception/PR, regulation and lack of ability to attract the right developers or employees were highlighted as concerns by practitioners.

In the second study, a pilot study, we investigate perceptions of AI and how this affects opinions on AI technology. We identify differences in perception between those who work in development of AI and ML and those who do not. We then make conclusions on how the perception of AI or ML may be quite different to the reality of the system that the developer or user is interacting with. Within this study, we highlight any interesting observations pertaining to how a developer or user perceives the technology and whether this affects their attitude towards, or their use of, it. Although initially we planned to employ interviews, the research was conducted by online survey due to difficulties with respondent availability for interview in 2020–2021. The surveys within this study have been conducted across a broad range of industries to investigate the current challenges around fairness [9, 66] in ML and AI development as well as needs for further support. This survey had a larger volume of responses, and therefore we were able to extract themes within the data that might support any incongruence between the developers' attitudes and perceptions and those of the user, and/or public. To our knowledge, this is the first investigation of this type which uses multidisciplinary context (Sociology, Philosophy and Computer Science) to understand the human element behind the modelling.

By conducting the second study, we found that perception emerged as a contributory factor to challenges faced by the practitioner. Practitioner's perceptions of what the user would like and the reality of the end user perception sometimes did not match or left a technical gap. In trying to meet the expectations of the perceived end user [82, 87], the technical and regulatory aspects of model development, in some cases, became almost secondary. This left a potential technical gap which then could threaten the robust nature of the model development. Perception is a key element of human nature and the perception of a system by a user may not reflect the reality of the system after a marketing team has used anthropomorphism or other methods to advertise the end product [22, 27, 54, 76, 96, 101]. This can lead to such conclusions by the non-specialist such as 'the robots are taking over the world' or 'I will lose my job to an algorithm'. Perception issues could also lead developers to misunderstand what the system is capable of, what the user wants it to achieve and the importance of robust modelling. This can then lead to an inflated sense of system capability.

Based on the findings from these studies, we highlight opportunities for practitioners, research communities, society and industry to work in an interdisciplinary manner. This would lead to a positive impact on driving forward development in modelling methodology in AI and ML. The findings are presented in a visual format for direct comparison purposes and no statistical testing has been used. In addition, COVID caused the survey to be done online instead of in person. No alterations were made to the survey in light of this.

Through the investigation, we identify challenges with lack of relevant tools, lack of industry specific tools, access to talent and budget as well as challenges with regulation and the industry necessity for good PR [14]. As with the results of Holstein et al. [44], we find again that literature continues to focus on 'de-biasing' whereas our respondents were more concerned with data collection and meeting regulation and/or avoiding PR disasters. This indicates a continuing discrepancy between research literature and real-world requirements. In agreement with Holstein et al. [44], we find again that practitioners continue to struggle with the application of regulation. What fairness and bias means to the public seems, in some cases, to be more important than the regulation itself. For example, if the issues that come with a hastily implemented model can be glossed over with PR, the attitude seemed to be to save money on due process and then pay for PR if there is an issue. Indeed, where practitioners were lost in a sea of legislation and regulation, it could seem easier to pay for PR than try to find the single source of truth to adhere to. Increasingly, ethics was quoted in the study as being an emerging, but critical, topic for societal acceptability of model implementation and development. In conclusion, recommendations are made for ways forward.

In the next section, the relevant background and related work to our two studies are presented and discussed. We then present the two studies and the results. We then conclude the paper and outline directions for future research.

2 Background and related work

Despite widespread attention to concepts such as biases and unfairness in ML and AI, to the best of our knowledge, only two prior studies, by Veale et al. [94] and Holstein et al. [44], exist that are similar to this study. Holstein et al. [44] conducted a series of semi-structured, one-on-one interviews with a total of 35 practitioners across 25 ML product teams from 10 major companies [44]. To investigate the prevalence

and generality of the key themes that emerged in those interviews, Holstein et al. [44] then conducted an anonymous survey with a broader sample of 267 industry ML practitioners [44]. In both studies, the authors uncovered several disconnects between the real-world challenges that arise in public-sector ML practise compared with those commonly presumed in the fairness ML literature [44].

Holstein et al. performed a study using semi-structured interviews with 267 Machine Learning (ML) practitioners. This was one of the first studies to highlight practitioner challenges and needs for support in developing more robust models. The study raised multiple areas of interest:

- Communication between model developers and data collectors
- Guidance on data collection
- Support in identifying data
- Guidance on bias within developers and the model
- Need for a more proactive, holistic and domain-specific auditing process
- Guidance on how to prioritise fairness[2]
- Understanding the impact of data on the model and fairness
- Support in addressing detected issues and avoiding any side effects.

Despite the many issues raised by the Holstein et al.'s study [44], we have not yet seen a substantial move forward on these issues. We currently see a focus on building more models to try to check or determine biases within existing models. This seems counterintuitive when many practitioners have complained that they cannot initially obtain the correct data for their modelling purposes [1, 50]. In the following sub-sections, themes are explored that could contribute to the continued misperceptions surrounding AI; precision, perception, trust and communication (media).

2.1 Precision

Veale et al. [94] investigated the challenges of ML practitioners. Veale et al. [94] conducted interviews with public-sector ML practitioners working across a range of contexts such as predictive policing [7, 25, 52, 58] and child mistreatment detection [18], to understand the challenges faced by practitioners in constructing ML systems that aligned with public values. In the study conducted by Veale et al. [94], the main identified challenges were:

- Users not understanding the modelling output
- Lack of buy-in for the model
- Lack of interest if the modelling was not done to the stakeholders preferred metrics
- Inability to augment the model with context or additional knowledge
- Lack of ability to scale the model
- Lack of detailed explanations of how the models worked
- Lack of understanding of performance and precision of models by stakeholders
- Resourcing problems and single points of failure
- Over reliance on an algorithm.

The findings of Veale et al. [94] concern language use [20] in relation to ML and the understanding and correct use of modelling methodology [92] by both practitioners, users and stakeholders. Similar challenges were found again in the later 2019 Holstein et al.'s study [44] and later again within this 2021 study (i.e. Study 1 reported in this paper).

In both studies, the authors uncovered disconnects between the real-world challenges that arise in public-sector algorithm development compared with those commonly cited as being the most prevalent in present ML literature. Despite the multitude of issues raised by the Holstein et al.'s study [44], such as on data collection, model building and development, as well as fairness and bias, we have not yet seen a substantial move forward in this area. We currently see a focus on building more models to try to check or determine biases within existing models [1, 50].

2.2 Perception

One large driver of many of the issues seen by Veale at al. [94] could be perception. Practitioners may understand the way in which the model works and understand the technical descriptions of precision and accuracy, but the stakeholders may not. Language and expression of technology have become a barrier between practitioners and users as illustrated in the example by Veale et al. [94]: "We have a huge accuracy in our collision risk, but that's also because we have 40 million records and thankfully very few of them crash, so it looks like we have 100% accuracy—which to the senior managers looks great, but really, we only have 20% precision. The only kind of communication I think people really want or get is if you say there is a 1/5 chance of an accident here tomorrow— that, they understand". In this breakdown of communication, we see stakeholders losing interest when they believe their metrics have not been examined and an overall lack of confidence in the model due to this lack of understanding. In addition to this, there is public perception, which is critical now that AI is being developed for such things as healthcare diagnosis and allocation of credit [12, 30, 85, 93], leading to the perception, or reality, that many areas of a person's life may now be controlled by

[2] A concept that is currently not defined in sufficient granularity to enable prioritisation of it.

an algorithm, sometimes with the perception that there is no right of appeal.

This is supported by further studies that have raised issues concerning how people perceive AI. For example, Shin et al. [86] state that despite the surging popularity of AI implementation, "little is known about the processes through which people perceive and make a sense of trust through algorithmic characteristics in a personalised algorithm system" [86]. Their findings "suggest that AI and future algorithms must transcend functional transparency or mechanical accuracy and fulfil actual user needs and requirements" [86]. This puts users at risk as we do not yet understand how they allocate trust to certain systems and whether this trust could be misplaced.

Samuel et al. [82], through a study of Artificial Intelligence researchers in Higher Education Institutions in the health faculty, found that the "interviewees viewed AI systems solely as a methodological instrument, one of a number of non-exceptionalist research tools. This contrasted with the media's portrayal of AI that optimistically focussed on the benefits of these tools" [82]. Samuel et al. [82] went on to state that a stakeholder can often be problematic when issues of scientific rigour are raised, and "may be less interested in hearing about the uncertainties of scientific practice" [82]. This reluctance to consider the ethics behind the model can have implications in terms of the responsible societal use of research [11, 61] and on the research and policy environment [72]. When users or managers cannot, or will not, engage with the workings of a model, it becomes difficult to ensure the construction has been ethical and robust. In addition, if they do not understand the model, they are at risk of being exploited by it.

Perception of a system is critical as argued by several researchers [3, 48, 65, 74]. Astington and Baird [4] describe how perception influences the language used to describe one's world. The perception of something as complex as Artificial Intelligence might very easily lead to language being used to describe it that causes misconceptions. Indeed, Crawford [23, 95] states that "We think of artificial intelligence as something floating above us, disembodied, suspended and without earthly costs or consequence". Crawford further states that "such imaginaries misdirect people from what is unfolding in the real world, the material world. AI is anything but immaterial; for its very existence, it relies on an earthly and unsustainable supply chain" [95].

2.3 Trust

Neri et al. [67] highlight the role of experts and state in the public's perception of AI, in that some experts were able to establish themselves as public commentators and create an idea that AI could be a real threat and endanger all humans. In this sense, Neri et al. posit that the experts framed and communicated a message that impacted public perception significantly. "This message of risk was based on counterfactual scenarios instead of actual events, such as any particular self-driving car crash. The counterfactual scenarios were at the basis of the messages of existential risks related to AI that were transmitted and amplified" [67]. Neri et al. [67] found that when pragmatic experts are forced to position themselves in public and take a stance themselves, they play a clarification role. Whilst they reject extremely speculative scenarios, some may want to stress the "real" dangers of the technology. This creates a new message. However, if pessimist experts this can trigger many indirect effects within society [67].

Indeed, Fast and Horvitz [31] have found that "discussion of AI has increased sharply since 2009, and that these discussions have been consistently more optimistic than pessimistic. However, when we examine specific concerns, we find that worries of loss of control of AI, ethical concerns for AI, and the negative impact of AI on work have grown in recent years". AI is a specialist area and a complex abstract concept, so there are barriers to the general public understanding what it is and how it works, and this can create perceptions of AI that are not factual [31].

Zhai et al. [100] conducted a study on the public's perception of AI in relation to the media. The findings were that "different subjects are competing for and dividing up the right to speak of AI, leading to the gradual fragmentation of the concept of AI. Second, reporting on AI often includes reference to commercial institutions and scientists, showing a successful integration of science and business. Moreover, the result of their topic modelling shows that news media mainly defines AI from three perspectives: an imagination, a commercial product and a field of scientific research" [100]. The prevalence of 'experts' speaking on topics such as AI and trust has led to an almost bandwagon effect. Ethics boards have sprung up, eager to make a name by providing reverse engineered ethics to model development. 'Experts' speak globally on their understanding of the problem without any significant training in AI or model development. This leads to users being seduced by marketing and anthropomorphic devices such as Siri and ascribing trust to these devices as well as chatbots and internet sites that is unwarranted [95].

2.4 Communication

It is difficult to adequately convey or describe abstract concepts as stated by Hayes and Kraemer—another "sense in which the term abstract is often used denotes a concept lacking a tangible referent in the real world" [42]. Yao et al. [99] describe a plethora of issues in describing concepts such as size, in our case we have to describe something that is

indicated as being 'as we are' as humans but yet built in metal and existing in the cloud. This becomes a concept increasingly more difficult to convey in language. Therefore, Zhai et al. [100] and Crawford [23, 95] are right to be concerned with how AI is portrayed and how the public perceive it. This could influence behaviours, attitudes and opinions of this difficult to access concept. This indicates the importance of perception, and indeed language, around complex technological concepts such as AI and ML.

In the next section, we describe the methodology used for our two studies: Study 1, focussed on AI/ML practitioners and practises around fairness, and Study 2, focussed on AI perception.

3 Study 1: needs of practitioners

This section discusses the methodology used within each survey. Both studies went through ethical review at the University of Portsmouth and were approved. The surveys are provided in full in the supplementary materials.

3.1 Methodology

For Study 1, the aim was to re-run the Holstein et al.'s survey to determine what, if any, progress had been made since 2019 on the issues raised by practitioners. The respondent cohort for this was practitioners within the fields of AI, ML and Data Science or indeed anyone working in these disciplines. Practitioners were asked the same questions as Holstein et al. asked in their initial study. Anonymous online surveys using JISC online surveys were conducted. Emailing lists were also employed. Additionally, the survey was sent to the group 'Women Leading in AI[3] but there was no response. The surveys went out on the Sprite+ Network[4] and were dropped into the chat box at a TAS Conference.[5] The surveys were also dropped into the chat box on international networking meetings. The surveys were announced on social media to approximately 2000 people plus further connections and several online communities related to ML and AI on LinkedIn. The surveys were also sent to special interest groups as well as networking groups on Slack. The surveys were promoted on Instagram to over 100 people. The Institute for Science and Technology[6] also sent the surveys out in their newsletter. These studies targeted the same practitioner types as the Holstein et al.'s study.

We advertised to over 2000 people and directly approached over 100. Twenty-one people responded to this survey. The survey data were compared to the previous Holstein et al.'s study in raw form. Therefore, no statistical testing was performed on the data.

An interesting pattern emerged that some practitioners did not feel to take the practitioner survey but were happy to take the more general user survey and answer the practitioner-related questions within the user survey (from Study 1). Feedback from practitioners was that their company may not look favourably on them answering this particular survey. This is because some answers could appear to show the company in a negative light. This is similar to the issues found in the Holstein et al.'s study "our contacts often expressed strong fears that their team or company's identity might leak to the popular press, harming their reputation" [44].

We notice that there is a clear diversity imbalance for both surveys. We disseminated the survey to many groups; however, a majority of white males answered. This means that threats to validity include a narrow cross-section of society that does not include other genders or backgrounds and lacks diversity in the responses. This is the view of a particular section of society and may not represent the whole picture. Therefore, the answers are not generally applicable but do give some, if limited, insight into practitioner attitudes since 2019.

The survey was designed with multiple choice answers and open text fields for additional context and for a respondent to specify if none of the above options applied and why. Open-ended questions were used to gather any additional information the respondent felt important to share. Due to branching logic, some survey questions were completed by a subset of respondents. In these cases, question response rates are provided in addition to the percentage of respondents who were asked the question. To illustrate general themes, we share free text responses.

In using calls for participation, we may have sampled practitioners who are motivated to discuss and address fairness issues in their products; however, even within this sample, it is noted that many teams are still reporting challenges to incorporating fairness into their products. The subsequent discussion focuses on technical and non-technical problems that may be contributing to this.

4 Results and discussion of Study 1: needs of practitioners—what has changed?

In this section, the results are compared to the study by Holstein et al. [44]. We address first, the differences in demographic and then we proceed with the following subheadings for analysis and comparison to Holstein et al.:

- Fairness aware data collection

[3] https://womenleadinginai.org/.

[4] https://spritehub.org/.

[5] https://www.tas.ac.uk/.

[6] https://istonline.org.uk/.

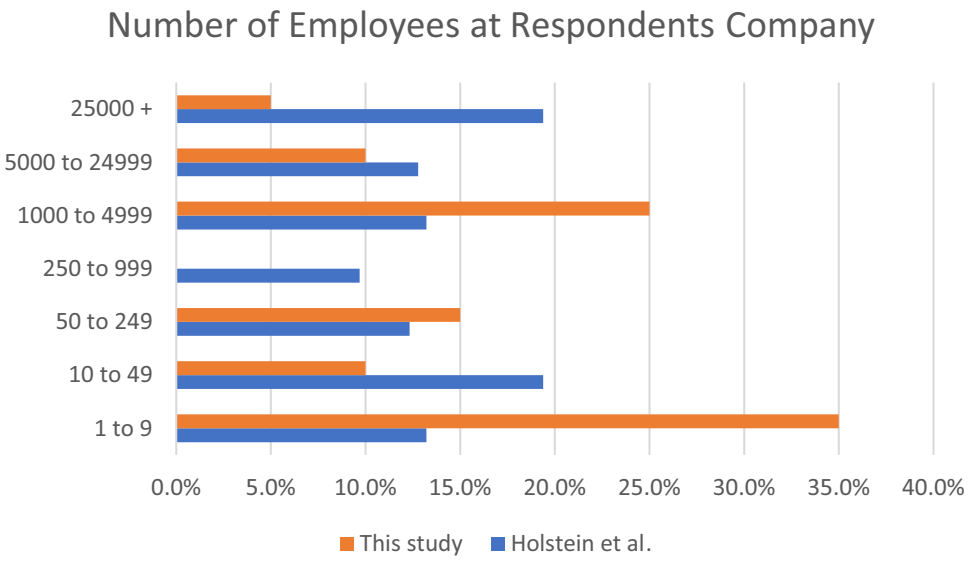

Fig. 1 Comparison of number of employees at respondents' company comparison

- Challenges due to blind spots
- Needs for more proactive auditing process
- Addressing detected issues
- Other factors
- Limits on AI being implemented

4.1 Survey demographics

Respondents were 50% from SME size companies under 250 employees and 50% from larger companies of between 1000 and 25000 employees (Fig. 1).

In comparison to the Holstein et al.'s survey, our survey reached predominantly small- and medium-sized enterprises with a good spread up to the 24,999-employee bracket. Where this survey differs is the level of engagement, we were able to achieve with very large businesses compared to Holstein et al. who had significant engagement from this sector. However, this study was able to gain a better insight into the small- and medium-size business—which is one of the most impacted sectors with new reforms and legislation.

4.2 Roles

The declared roles by all respondents are displayed in the results below. Many additional roles were declared, with some by the same respondent undertaking multiple roles within a company or project. Some did not choose a specific role from the list provided and declared other roles as displayed in the below results. The top reported team roles by respondents in both studies were Data Scientist and Researcher. The study by Holstein predominantly engaged with the Data Scientist and Researcher role (Table 1). This might be due to a change in titles for employees over the years and this might account for why this study had a lot more employees declare themselves in the 'other' category (Fig. 2).

We found that the cross-section of roles obtained from our study corresponded to that of Holstein et al. However, due to participants declaring more than one role in most circumstances, it is difficult to know what the immediate challenges of the role would be. For example, if a director of a small firm is also the software engineer and researcher, they may not be qualified in every area and so may have different challenges to those who have significant experience or qualification in that role. The naming conventions of roles change periodically, so we added in the further list of unique responses where participants declared their role as 'other'. The results are below.

4.3 Domains and technology areas

The domains that the teams work in are varied as seen in the following graphics. Some felt that they could not choose from the available options and declared additional domains. Some also declared additional technology areas. The top declared technology areas in this study were healthcare and

Table 1 Additional declared roles

Additional roles in Holstein et al.	Additional role in this paper
Head of data science	Project collaborator
Machine learning engineer	Business analyst
Data curator	Sales
CEO	Citizen researcher
Scientist	
Research development	
Innovation manager	

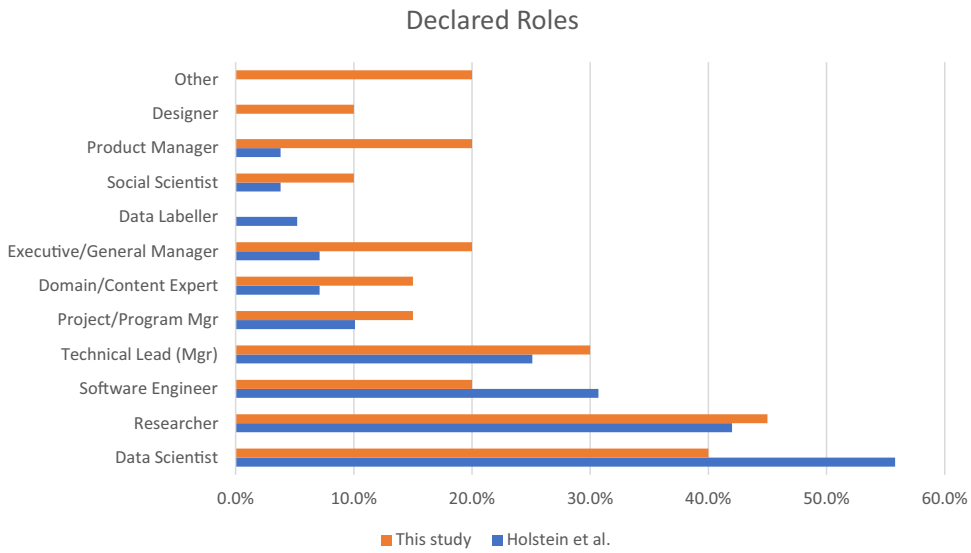

Fig. 2 Declared roles' comparison

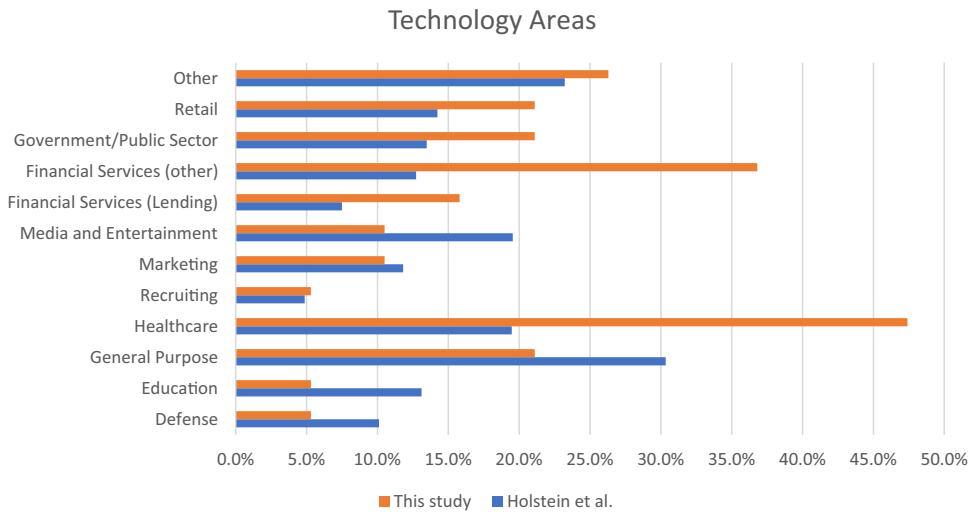

Fig. 3 Technology areas' comparison

financial services vs. the Holstein study results of General Purpose and Other. One clear increase was in healthcare; given the issues with COVID, this is a reasonable change in employee industry to observe. This survey also engaged more with the financial services, and 'Other' category (Fig. 3).

The application domain also varied greatly as seen in the previous two figures. The top declared domains in this study were predictive analytics and decision support as compared to those of the Holstein study which were natural language processing and predictive analytics. Again this study engaged more with 'other' which might be showing a changing of application domain over the years or simply that our target audience was made up or more 'other' (Fig. 4).

The types of roles might not have been captured fully so the option was given to input a specific role of the participant felt they needed to. The list of unique responses where participants declared their role as 'other' is given below (Table2).

Participants spanned a wide range of roles, companies and contextual areas. However, many common traits were observed. In the following analysis we discuss the common challenges and needs around fairness organised by high level themes along the same lines as the study by Holstein et al. [44] for comparison purposes. We then examine sub-themes of research and design opportunities.

4.4 Fairness aware data collection

This section examines what challenges are associated with initial and subsequent data collection. The methodology of data collection and consultation of relevant experts is also explored.

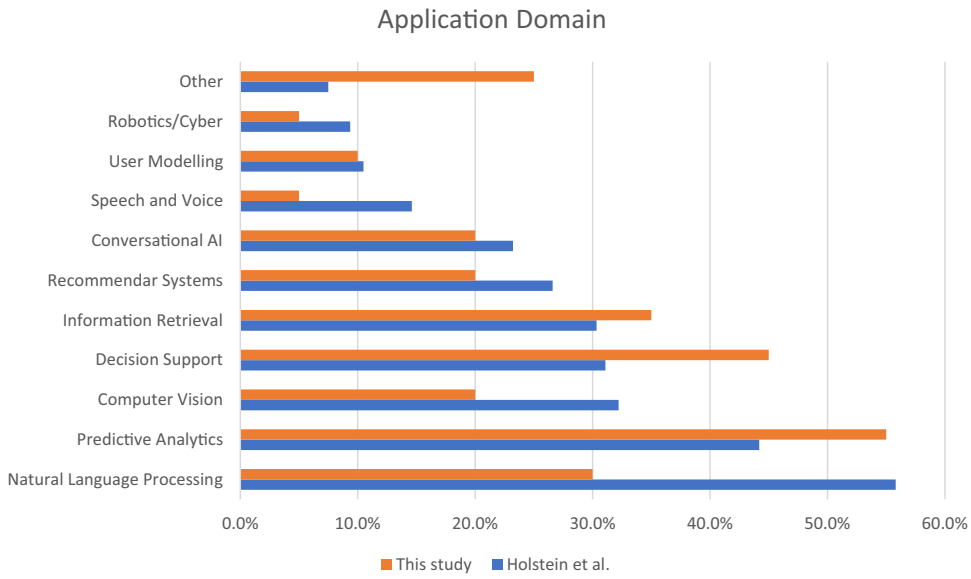

Fig. 4 Application domain comparison

Table 2 Additionally declared domain applications

Domain	
Holstein et al.	This study
Energy	Customer services
Natural resources	Legal
SaaS	Manufacturing
Sales	Robotics

- 45% of respondents reported prioritising fairness in their work, whilst 40% prioritised this moderately or a little and 10% did not prioritise this at all. Interestingly 40% have just begun prioritising fairness with an additional 10% starting to prioritise in the last 6 months.
- Only 15% started to prioritise fairness over 5 years ago, the rest being more recent.

Reasons for prioritising fairness are then discussed as follows:

- 40% of all respondents reported this was due to a desire to show the public they were prioritising fairness.
- 50% of all respondents reported that they wished to avoid being perceived as unfair by customers.
- Overwhelmingly 68.4% responded that they did not wish to violate legal requirements on fairness in AI.
- 57.9% responded that they wished to avoid a potential PR disaster.
- 60% reported prioritising fairness due to a sense of ethical responsibility [60] with 95% of all respondents stating a desire to improve the quality of their products and services by considering fairness in development.

Free text comments on this area include:

(1) "Regulators, particularly around PII have and the ability to explain the model are fair are increasingly impacting our work".
(2) "Mission from inception has been to overcome complex, difficult problems, not to make matters worse."
(3) "A core part of our business is credit scoring, which is fundamentally about fairness - it's not an add-on".

4.5 Comparison of Holstein et al.'s [44] study

In the following, we compare the results of our study with ones from the 2019 study by Holstein et al. [44].

Holstein et al. [44] reported that "Out of the 65% of survey respondents who reported that their teams have some control over data collection and curation, a majority (58%) reported that they currently consider fairness at these stages of their ML development pipeline. Furthermore, out of the 21% of respondents whose teams had previously tried to address fairness issues found in their products, the most commonly attempted strategy (73%) was "collecting more training data"" [44]. We show a comparison in the following table (Fig. 5).

This shows an increase in practitioners who have control over the data but a decrease in those who consider fairness at this stage of the development pipeline.

In Holstein et al. [44], "52% of the respondents who were asked the question (79%) indicated that tools to facilitate communication between model developers and data collectors would be "Very" or "Extremely" useful." In this study, this was mirrored by 60% of respondents indicating the same wish.

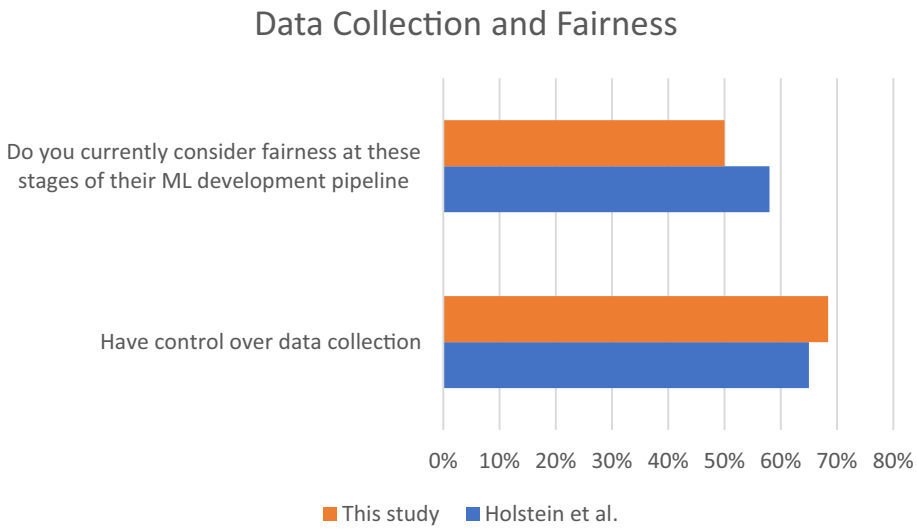

Fig. 5 Data collection and fairness considerations

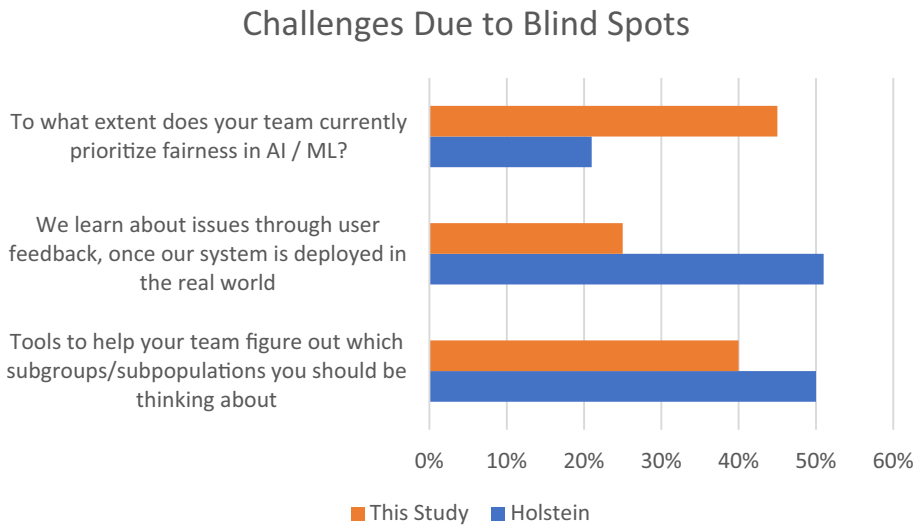

Fig. 6 Fairness

In Holstein et al. [44], it was stated that: "To score African American students fairly, they need examples of African American students scoring highly. But in the data [the data collection team] collect[s], this is very rare. So, what is the right way to sample [high scorers] without having to score all the essays? [...] So [we need] some kind of way... to indicate [which schools] to collect from [...] or what to bother spending the extra money to score."

In the current study, the following is commented: "Balancing data training sets to include the equal size of possible sets looking from gender, age or race. It helps to create a more consistent and sustainable models that are not dependent on some dominant feature. For example, the female dominates the data set, it's not necessary to create balanced data sets for training because models will over train for dominant ones and start ignoring others. It is a more technical issue, not only ethical" (respondent id 76282410).

This indicates a continuing need for methodologies to ensure fair data sets.

4.6 Challenges due to blind spots

In comparison with Holstein et al., we see in the following table, less respondents reported finding issues with their system when deployed in the real world through user feedback. A similar percentage of respondents to the Holstein study would like tools to support fairness aware data collection. And more respondents in this study than the Holstein study prioritised fairness. This could indicate a general move within the practitioner community to prioritise fairness. However, as we saw earlier, the reasons for prioritising fairness are predominantly to avoid litigation and PR disasters, both of which could seriously damage a company. Therefore,

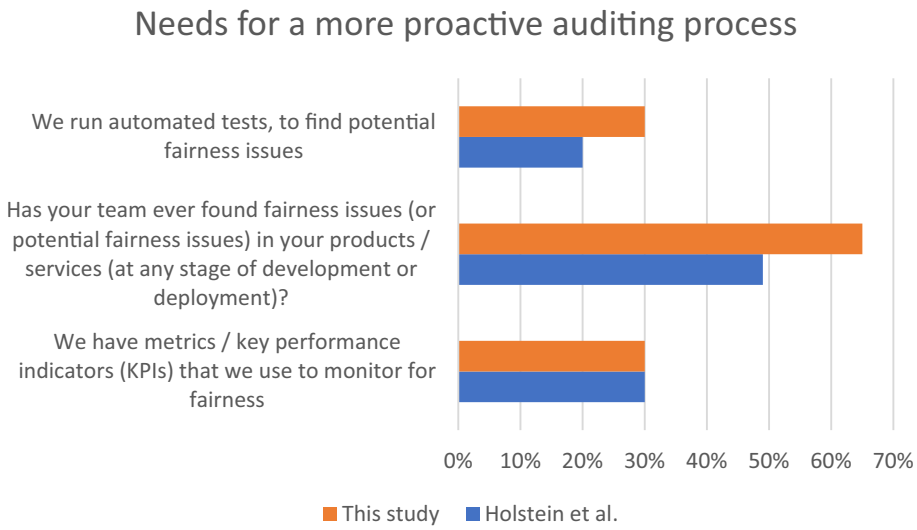

Fig. 7 Needs for more proactive auditing process

it is not surprising that companies are now starting to prioritise fairness in modelling (Fig. 6).

4.7 Needs for more proactive auditing processes

A comparison of the Holstein study vs. this study is show in the following table.

The same level of respondents have key metrics to monitor fairness in development, but this is still very low at 30%. More respondents in this study are finding issues either during development or during deployment. This is a significantly higher percentage than in the Holstein et al.'s study. This indicates that little progress has been made in this area. It might be that board involvement would be required to construct key metrics to work with and this may have not been forthcoming due to a multitude of reasons. Reasons could consist of, too much guidance, unsure what guidance to base indicators on or a lack of resource to call the meetings required to establish KPIs (Fig. 7).

4.8 Addressing detected issues

There are significant discrepancies between the Holstein et al.'s study and this study in terms of addressing detected issues. 71% in Holstein et al stated that it would be very useful to have tools to understand side effects of fixes within the model, whereas in this study, only 11.2% thought this might be useful. This is interesting as it either tools to determine the side effects of 'fixes' have been established and implemented or that this is no longer the biggest concern of practitioners. Furthermore, if tools have been established for the other three requests, then maybe the tool for determining the side effect of 'fixes' is no longer required. Conversely, 65% of respondents in this study then requested tools to help with deciding on population and sub-population data. Less respondents than in the Holstein et al.'s study thought that tools to navigate ethical choices would help. The same percentage of respondents as in the Holstein et al.'s study would like tools to reduce human biases. It is clear that tools in all the categories below are still being requested but are clearly not forthcoming. In the next section, issues with the volume of regulation are cited and this might be holding back the willingness or ability to develop tools to address such issues as fairness and ethics, these two terms being difficult to define and with no agreed definition. It seems that there is still an ongoing request from the majority of the practitioner population for some way to address detected issues (Fig. 8).

4.9 Other factors contributing to fairness implementation

Other factors contributing to fairness implementation can be seen in the Table 3. These are free text responses from this study. The main issue appears to be training of the model and collection/use of the data. Resource is also specified as a challenge to being able to address fairness related issues in model development.

4.10 Limits to AI being implemented

Two responses were recorded to the question of what might limit AI being implemented. Regulation is cited as being too plentiful and hard to keep track of. It is also cited as being difficult to implement due to the volume and pace of development. This is also indicated in the fact that some practitioners were unaware of some key regulation in Study 2.

Additionally, in Table 4, respondents reported additional limits on AI being implemented in their business related to issues with regulation and potentially the difficulty in understanding and adhering to it.

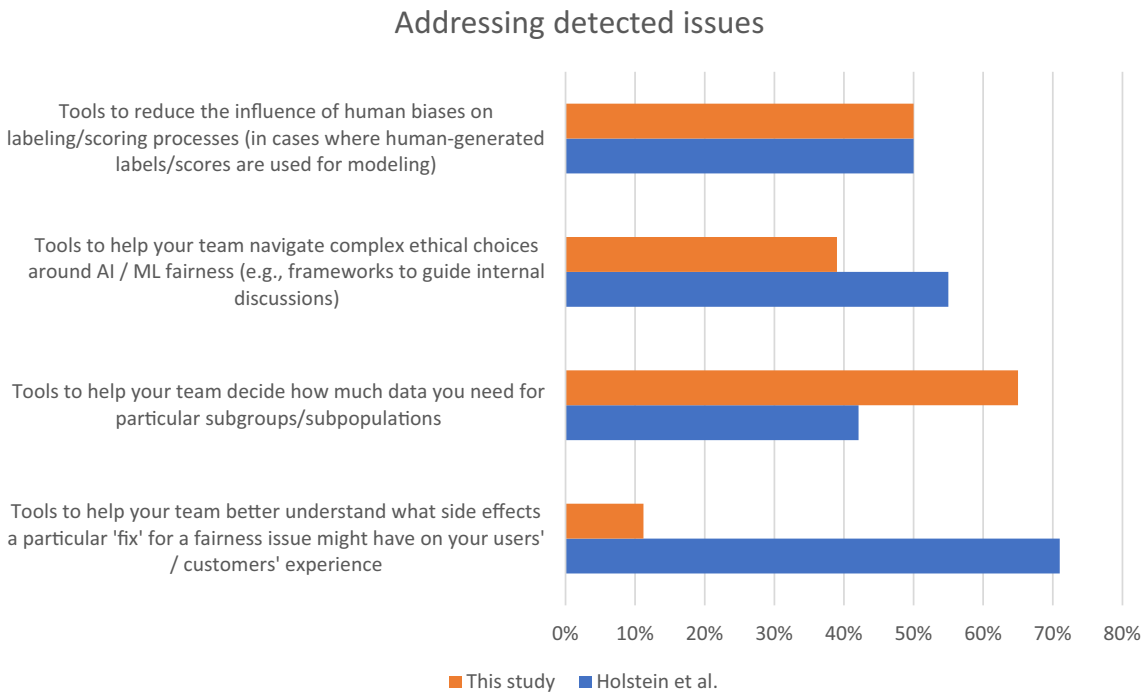

Fig. 8 Addressing detected issues

Table 3 Other factors contributing to fairness implementation

The RL setup if different to classical model train then tests approaches. The aim is for a model to learn from its environment in an adaptive manner. The application can be quite good context specific and may be biased in a statistical sense for very good reason
Considering and implementing variants in perspectives amongst entity users throughout the field
We have a small team and found it difficult to attract qualifies members without the funding of big tech
Main issues with AI/ML are detection of unusual patterns (bias) and subsequently understanding of how these have developed. How can they be found so that the system is re-optimised to be fair and safe
We can collect all data but are selective in what we include in our models

Table 4 What might limit AI being implemented in respondents' business

Regulatory compliance. Unexpected outcomes form the information even after validation so not meeting business expectations
Regulation

5 Study 2: pilot—perceptions of AI

Study 2 is a pilot study combining technical- and perception-related questions; thereby combining the fields of Philosophy, Psychology and Computer Science. This study has been undertaken to gather some introductory data to inform a larger study in this area. We investigate perceptions of AI and how this affects opinions on AI technology. We identify differences in perception between those who work in development of AI and ML and those who do not. We then make conclusions on how the perception of AI or ML may be quite different to the reality of the system that the developer or user is interacting with. Within this study, we highlight any interesting observations pertaining to how a developer or user perceives the technology and whether this affects their attitude towards, or their use of, it. The research gap here is that there has been no progression on perception of AI in relation to practitioners since the study by Veale et al. [94]. This is the only study of its kind as far as we are aware.

5.1 Methodology

The respondent cohort for this was practitioners developing, or users of, models within the fields of AI, ML and Data Science. Anonymous online surveys using JISC[7] online surveys were conducted. Emailing lists were also employed. Additionally, the survey was sent to the group Women Leading in AI[8] but there was no response. The surveys went out on the Sprite+ Network[9] and were dropped into the chat box at a TAS Conference.[10] The surveys were also dropped into the chat box on international networking meetings. The surveys were announced on social media to approximately 2000 people plus further connections and several online communities related to ML and AI on LinkedIn. The surveys were also sent to special interest groups as well as networking groups on Slack. The surveys were promoted on Instagram to over 100 people. The Institute for Science and Technology[11] also sent the surveys out in their newsletter.

We advertised to over 2000 people and directly approached over 100. One hundred and one people responded to this survey. $N > 20$ is a statistically valid sample, and so the analysis was processed. The survey data are displayed in raw form. Therefore, no statistical testing was performed on the data.

A clear diversity imbalance was noticed for both surveys. We disseminated the survey to many groups; however, the largest respondent category was white, middle aged male.

The survey was designed with multiple choice answers and open text fields for additional context and for a respondent to specify if none of the above options applied and why. Open-ended questions were used to gather any additional information the respondent felt important to share. Due to branching logic, some survey questions were completed by a subset of respondents. In these cases, question response rates are provided in addition to the percentage of respondents who were asked the question. To illustrate general themes, we share free text responses.

The survey design was multiple choice with open text fields so that respondents could add extra context or information they felt relevant.

Due to branching logic, some survey questions were completed by a subset of respondents. In these cases, question response rates are provided in addition to the percentage of respondents who were asked the question. To illustrate general themes, we share free text responses.

As stated previously, many of those that felt they were not able to answer the first survey did go on to answer the second and went on to answer similar questions to survey one that we had included via branching logic. Feedback from practitioners was that their company may not look favourably on them answering the first survey. This is because some answers could appear to show the company in a negative light. The aim of this survey being different might have attracted the respondents to answer.

6 Results and discussion of Study 2: perceptions of AI

In this section, we present the results of Study 2, including demographics, views on automation, fairness and use of technology. We present results of practitioners who refused participation in Study 1 but agreed to participate in Study 2.

To determine the attitude of users to technology, we wanted to know what their views on automation and fairness were as well as their opinions on their own use of technology as to whether they found it effective or not. An additional section asked practitioners some of the same questions as in the first study. This was included on purpose to see if we could get more responses under a different survey title. Whilst reaching out to specialists and practitioners, we found many were concerned about how their participation in survey one would look to their company, but they had no concerns over answering survey two.

6.1 Demographics

This survey has a sample size of 101 people of which 72% were male. The age of respondents was varied, with the majority of respondents being in the age group 21–50. This produced a bias sample towards educated white males. This shows a lack of diversity.

The breakdown is as follows:

- 70% were white British
- 77% had degrees or higher degrees
- 78% were involved in technical professions.

6.2 Views on automation

- 75% of respondents thought that automation either could not happen in their role or that automation of their role would happen after retirement; unsurprisingly, 83% did not feel worried about this.
- 70% of respondents were unaware of, or did not know about, the ICO Guidance on AI and explainable AI[12] [27,

[7] The online survey tool designed for Academic Research, Education and Public Sector organisations.

[8] https://womenleadinginai.org/.

[9] https://spritehub.org/.

[10] https://www.tas.ac.uk/.

[11] https://istonline.org.uk/.

[12] Information Commissioners Office: https://ico.org.uk/.

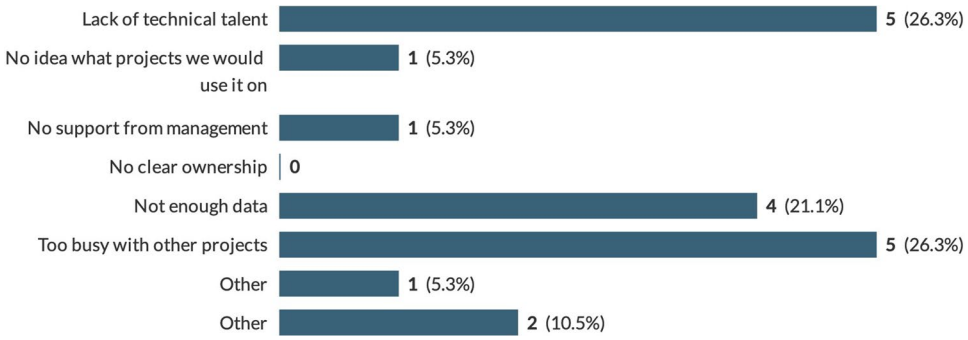

Fig. 9 What might limit AI being implemented in respondents business

29, 53, 69] or how this might impact their work. Of 28% of respondents that were aware of the ICO Guidance on Explainable AI, only 11% thought that this guidance was adequate. 86% were unaware of any other guidance or guidelines on explainable AI. In support of this, 60% thought that it should be law that an AI system should have to explain a decision made about them with a further 32% stating this would be good practise.

- This indicates a discrepancy between people asking for regulation at the user end but being unable to provide it during development. The explanation of decisions will only become a reality if the challenges in this paper are addressed.
- The most common cited barriers to AI being implemented within a business were 'too busy with other projects' (26%), 'not enough data' (21%) and 'lack of technical talent' (26%). This reflects what was found in the first study and in Holstein et al. where data barriers were cited as significant for those developing models. Lack of technical talent is also reflected in the first study as being a critical issue. The disconnect between what industry wants and what the practitioner can deliver in the new technological areas of Data Science, Machine Learning and AI is one that is critical to close to ensure correct hiring practises (Fig. 9).
- 57% felt neither confident nor un-confident in using AI or ML products. This is an interesting result, and it was not possible to pinpoint a reason as to why the result of this question was predominately neutral. 15% felt confident and 26% did not feel confident.

6.3 Fairness

- Of the 19% that answered that they worked on ML and AI products, 63% had control over the data collection process and 73% felt that they considered this well. However, 68% did not believe that they had enough data for the modelling process, but 56% felt that they had data fit for purpose.
- When developing products, 84% considered fairness and 47% had discovered fairness issues in their products. Only 36% added that they had an audit process for their model or product.
- 68% felt that they could validate and verify their model adequately and 84% stated that they knew what validating and verifying a model meant and were aware of the consequences.
- 68% did use caveats and assumptions to bind their model. 57% used cases to guide model development.

All of these results were more positive than the first study. This is an interesting result and one that might have been caused by the high degree of respondents that rated AI in a positive light (78 respondents out of 101).

6.4 Use of technology

- On the subject of chatbots, 75% of respondents did not find then to be human-like and 82% would find a human on the other end of the phone more effective.
- 82% did not think that chatbots solved their problem first time. 73% stated that they had to bypass a chatbot to go through to a call centre.
- 34% did not think a voice recognition service worked for them with 64% stating it worked often, very often or always.
- When speaking about AI in particular, 18% stated that is overhyped with 65% thinking it would change their business to some extent. 54% were optimistic about this change.
- 63% also thought that AI would change their home life.
- Moving onto more philosophical questions, 46% thought that consciousness was uniquely human, with 28% stating that they did not know. Forty percent thought that computers could never be conscious, with 28% stating that they did not know.

From these results, we can see that although there is some trepidation concerning the effectiveness of new technology; overall, the attitudes are positive towards this change. The belief that AI would not affect their job but might affect their home was also seen as positive by respondents. The majority

Table 5 Free text responses

Comment	Respondent ID
The whole idea and prospect of AI and our future is exciting. Just like Isaac Asimov said hypothesised in his books, the next chapter of human is to put our consciousness into robots and then get the robots to go into space as they do not require any food, oxygen, do not become unwell, etc. In my opinion, AI is something that can change our lives for the better in a good way. I hear the concerns around AI but am not concerned but more excited about the prospect of new ideas!	652,267-652,258-67,088,458
I think we're 10 years away from viable companion robots for senior citizens	652,267-652,258-68,079,878
Some applications of what would be considered "AI" were absolutely excellent. Some applications of "AI" would worry me. At the same time, most of the applications of "AI" that I have seen thus far have been mundane, and it is relatively understandable how and why people try to do them	652,267-652,258-68,155,136
AI/ML is no where near real intelligence, it is merely pattern matching at a low level compared to human ability. True intelligence will come when computers can match the number of connections that a human brain has	652,267-652,258-68,172,143
AI will revolutionise the way we live, work and play—within our lifetimes. The potential benefits massively outweigh the risks	652,267-652,258-68,202,255
Chatbots I have encountered have been unable to deal with my queries, because my query has been something unusual, outside the range of scenarios the developer seems to have imagined is possible. This leads to greater frustration, because, first, I try the chatbot, then I need to hunt hard to find a phone number, and then I listen to the supplier's choice of music for a long time before talking to a person who can help. My concern with AI in general is that there needs to be something in the economy that causes people to have jobs to do to deal with the basic human need for meaningful activity. Also, I am concerned that AI could be used as a population control and manipulation tool leading to people having and experience of life like the medieval serfs who did not have the freedom to pursue what they believe to be good, and this will reduce the majority of people to a state of dependence on the good will of their 'betters'. AI used in limited applications, such as autonomous vehicles, might be beneficial so long as vulnerabilities to cyber security threats are overcome, and also privacy matters must be addressed which in turn might lose some of the potential benefits such as optimising the community travel through knowing where everyone is going as a method to avoid congestion delays	652,267-652,258-68,356,757
I believe that Elon Musk has a valid point, and it is the unethical development of AI that worries me. That's places outside the UK and will impact severely on national security	652,267-652,258-70,864,157
Human Intelligence and Consciousness (wisdom making ability) is always better than Artificial Intelligence	652,267-652,258-71,129,734
If a computer can theoretically simulate a brain, then I think it can achieve consciousness. What's the difference?	652,267-652,258-70,975,595
As AI advances, I think it can help a lot of people save time, money and effort. However, I predict there will be a plateau that we will all reach where we might become nostalgic or missing the human connection	652,267-652,258-75,885,754
I cannot emphasise enough that computers will never achieve consciousness. Technology can come close to give the illusion that a piece of tech is a sentient entity, but that is once again nothing but an illusion	652,267-652,258-76,652,969
The role of AI will move society forward. Some roles will be 100% AI, and others will be less affected. That's evolution and should not be challenged. We are biological cognitive machines, and we mimic ourselves in the AI	652,267-652,258-76,624,615

of respondents believe that either consciousness is uniquely human, or they do not know and the majority stated that they thought computers could never be conscious or they did not know. However, on the opposing side, it is interesting to note that 54% either did not know or thought consciousness was not uniquely human. This question was not expanded to define consciousness or include animals or computers, so respondents had a chance to frame their own understanding.

Below, we see free text responses to this survey (Table 5).

The above shows very different opinions on the future of AI. Some believe it is hype, some believe it will have an effect and others believe it will have an important place within our lives, perhaps a critical place. This disconnect of understanding of the technology and its capability shows why there may be such a discrepancy in the way practitioners approach development. This gap must be solved by education. However, technology is developing so quickly that it is has been difficult for education or the government to keep up with it [102]. The more optimistic may not believe the technology needs such oversight but the more pessimistic might. This could lead to less oversight in technological development, which in turn can lead to disastrous consequences for society, as stated earlier.

7 Isolation of results of practitioners who refused Study 1 but participated in Study 2

In an unexpected event in the data collection, some practitioners refused to complete the practitioner survey (perhaps for reasons already mentioned) but went on to complete the perception survey. This sub-section is treated as a pilot study, and despite not containing a statistically valid sample size, there are some interesting highlights that could be taken forward for future research.

The second study allowed practitioners, through branching logic, to answer some of the same questions as the first study. Therefore, in the user survey, we were able to draw some tentative conclusions on perception and AI implementation. Here, we discuss what this initial analysis illustrates about practitioners and perceptions.

In this study, 17 respondents declared that they worked on the development and/or implementation of ML/AI products, such as automated systems in finance [93], chatbots, predictive methods [7, 17, 89], healthcare [56, 82] and automation (Table 6).

8 Results

Of the 17 practitioners:

- 71% considerately modelled the data process.
- 35% did not know how to validate or verify a model, of which 17% did not know what the consequences would be if this was not undertaken.
- Of two respondents who answered that they understood validation and verification (V&V) [2, 36, 43, 59, 92], as well as the consequences of not undertaking V&V, they stated their business still did not do V&V.
- Despite answering that they understood V&V and implemented it, two did not know what caveats and assumptions were; 35% either did not know or did not use caveats and assumptions with their models.
- Seven of the seventeen respondents could recall times when they or their team have discovered fairness issues in their products.

8.1 Perception and AI development

By extracting respondents who answered that they were either optimistic or pessimistic about AI, the following was found:

- Of the 78 that were optimistic about AI, 35% were not positive about AI in their workplace; however, only 4 worked on AI or ML in this group.
- Of the 78 who were positive about AI, 17 or 21% worked in developing AI or ML.
- Of the 18 who were pessimistic about AI, 83% were pessimistic or agnostic about AI in their workplace with only 1 person working in the development of AI.

Those who work in AI seem to be more optimistic about the technology in general. Those who do not work in AI may tend to be more pessimistic in general. This could reveal an optimism bias in working in AI and a mistrust bias amongst those who do not work in the field.

9 Discussion

Without the contextual narrative that the semi-structured approach might have, given it is difficult to form conclusive opinions as to why the results are as they are. However, using recent studies and literature as well as the limited narrative gained from participant comments, we can form a picture around the issues.

In the first study, we can see that the issues we are observing are affecting mostly middle size and small businesses. The domain applications, roles and technology areas are all quite different in this study and this might be due to the participants we reached with this study but it might also be due to the changes that COVID had on the landscape of tech. Between Holstein et al. and this study, tech experienced a boom in popularity whilst people worked from home. Certain industries were forced to upgrade their tech quite quickly to cope with new scenarios and this might have been a cause for the larger number correspondents in healthcare and finance. Role names change quite often in the tech industry due to rapid developments so this might also account for the difference in respondents' roles.

This study shows that many challenges remain or have become even more prominent since the Holstein et al.'s study. Respondents report having slightly more control over data but less consideration of fairness. Respondents reported finding less issues within deployed technology, but they also prioritised fairness to a greater degree than those in Holstein et al. Nearly half still need tools to help with data collection. This might account for the recent policy changes that have led to requests for support by Chartered Statisticians in business critical and legal work. More respondents than in Holstein et al. found potential fairness issues through running automated tools to find them. Within the narrative comments range around understanding; "understanding [bias issues]", selective data use,

small teams and context specific models. This highlights current technical short fallings in those developing tech which is further exacerbated by inability to comply with regulation or not knowing which regulation to comply with. In trying to build working systems, most respondents were more concerned with avoiding a PR disaster or 'being seen to' comply with the guidance du jour. This was balanced by 95% of respondents wishing to do better in terms of ethical modelling. This indicates that the desire is there to do better but the finance, skills and profession are not. This is concerning given the time lapse between the Holstein et al.'s study and now as well as the surge in tech being implemented in the last decade. The analysis here highlights many of the issues already found in UK Government modelling and shows that they also exist to a lesser or greater degree in society. The uncertainty around robust and ethical modelling, what professions need to be involved, what guidance to follow and the basics on how to model are all impacting the future of this industry. The users and people who will be affected by the tech being developed could be impacted negatively by this.

As stated, some respondents did not participate in Study 1 but took the branch of Study 2 that asked the same questions as Study 1. The title of the second study seemed more amenable to these respondents and they were more likely to indicate a negative response such as lack of validation and verification process and knowledge in their business. Possibly as this survey was not seen as targeted at them or their business, the more technical questions were branched off from a study, more about perception. Those that were negative about AI in the workplace did not tend to work in AI. Only a fifth of those working in AI were positive about it. Of those declaring pessimism towards AI, 83% were pessimistic about AI in the workplace, but only one respondent worked in AI. This might show an optimism bias of those working within the field of AI and ML. Those that responded that did not work in AI tended to be more pessimistic.

Study 2 aimed to detect reasons for the negative effects of current implementation of technology within attitudes and perceptions rather than the technical aspects of the role. The results showed that many respondents were neutral about using AI and many did not want to, or could not, implement it anyway. 82% of respondents would prefer a human at the other end of the phone rather than using a chatbot. This shows that either humans prefer human connections or the technology that chatbots are based on is not yet good enough for deployment. Conversely, a large percentage of respondents were happy with voice recognition services. In the comments, those who were annoyed by voice recognition, chatbots indicated that they were particularly annoyed by them. This leads to an interesting aspect of emotion. The strength of emotion displayed may be that the expected function, when not performed, is extremely problematic for respondents. A small percentage of respondents believed AI was overhyped and a large percentage were optimistic about any future changes where AI would become more prevalent. This may be offset by the large proportion who believed this would not affect their role in their lifetime. Nearly half of all respondents believed that consciousness is uniquely human and that computers could never become conscious. This might account for some of the optimism displayed by the respondents. This is supported in the free text comments section where those that had negative opinion about AI were concerned about the control it could have over humans. Those who were positive quoted such aspects as saving money and time.

The over optimism towards AI might stem from the belief that no impact will be felt in the workplace and, possibly, the hype around AI and the increase in funding for AI applications might generate more optimism. However, this could be showing within the results of Study 1 where the optimism might be leading to lack of rigour in the modelling process and the use of PR as a sticking plaster to cover the gaps in the poor implementation and non-robust development of the tech. It appears that most practitioners would like to improve their robust modelling processes but either do not know how or the tools they are requesting are not forthcoming. What is clear is that without clear direction and guidance within the profession of modelling, we may not see much improvement, but there is a potential to see increased negative impacts on society.

Table 6 Other declared factors for prioritising fairness

Other	ID
A core part of our business is credit scoring which is fundamentally about fairness; it is not an add on	69,081,449
Our ML AI model building checklist is much broader and complex. It also provides as accurate information to both healthcare professional as well as their client	72,406,636
Balancing data training sets to include the equal size of possible sets looking for gender, age or race. It helps to create a more consistent and sustainable models that are not dependent on some dominant feature. For example, the female dominates the data set, and it is not necessary to create balanced data sets for training, because models will over train for dominant ones and start ignoring others. Its amore technical issue. Not only ethical	76,282,410
Regulators, particularly around PII, have the ability to explain the model are fair are increasingly impacting our work	76,347,398
Mission form inception has been to overcome complex, difficult problems, not to make matters worse	76,643,799

10 Conclusion

In conclusion, it appears that, overall, since the Holstein et al.'s study in 2019, very little has changed in the macro-view. Some aspects of practitioner challenge appear to have improved but overall, the call for more guidance and education on some aspects of modelling such as data, bias awareness and help with legislation and regulation remain.

We can see that there is a large disconnect between people's views of AI and its capability. This may affect the way practitioners develop technology and the way users perceive it. This relies on education to close this gap. However, as technology is progressing so rapidly is difficult for the education system, regulation or legislation to keep up with it. If we cannot understand the technology, or at least have some common view on it, then it is very difficult to create legislation or regulation around it.

The rapidly growing development and spread of ML and AI systems presents many new challenges. Automated systems are increasingly being used in broader and more varied industries, with ever more serious implications. We are entering unchartered territory that holds a vast array of consequences, some that we are yet to observe [70]. Therefore, we repeat the call of Holstein et al. [44] that "as research in this area progresses, it is urgent that research agendas be aligned with the challenges and needs of those who are affected by the technology". The recommendations outlined in this paper are opportunities for practitioners and research communities to become more robust and collaborative in the development and deployment of ML and AI systems.

11 Recommendations

As the world moves towards increasingly complex models without a common perception, language or understanding the need for fully robust modelling processes is clear. Increasingly concerns are raised around aspects such as privacy and fairness by practitioners and non-practitioners. Society has been disadvantaged over the last few years by inappropriate or non-robust modelling that has seen legal challenges launched against it and illustrates that many challenges remain. In this paper, we presented a study of practitioner's challenges and a study of user/practitioner perception. Together this, analysis illustrates a picture of confusion and competing priorities for practitioners, as well as confusion and potential distrust by users. Below, we highlight some directions for future research to reduce some of the confusion and mitigate some of the challenges. The main challenges have been taken from each study and recommendations have been formed below to help find a way forward.

11.1 Robust data collection: education and guidance

The survey responses indicated that collecting the right data in the right manner was difficult, with 68.4% reporting having control over the data. To facilitate more robust modelling, there must be a focus on supporting practitioners or bringing in the correct expert to collect and analyse high-quality data sets. This must be with a focus on how the data are collected and the statistical robustness of the initial data [41, 57, 91, 97].

- Understanding what data are available and how it relates to real world in collaboration with domain experts.
- Given the legal challenges over the last few years, we have seen that the real-life context of the data remains a challenge for practitioners. This has led to discriminatory algorithms due to a poor data set or a poor understanding of how the real world operates. This is a critical problem for practitioners who deploy algorithms into society [10, 15, 26, 33, 34, 40]. In this study, only 50% reported considering fairness at the development stage, in line with the other studies [28, 32, 38, 55, 64, 72, 83].
- Training should be implemented in this area from school level, so that the complex concepts such as assumptions, caveats, quality assurance and answering the right questions with constructive challenge become a cultural fixture [72]. This is also critical, because the students of today are entering a workforce and society that is more technologically oriented than ever before. To this end, we should train them to be practitioners of best practise. This can then be built on in successive generations [35, 75].

11.2 Audit

- Auditing was an area in which practitioners requested further support.
- Auditing well could potentially decrease legal challenges. This would provide a gateway for algorithms and models but also an opportunity for practitioners to consider such concepts as fairness, context and data before starting development of a model [17, 37, 68, 71, 72, 78, 80].

11.3 Ethics and bias

Further support on the inclusion of ethics and prevention of bias was sought by practitioners.

- Incorporating ethics into funding for AI could play a part in acting as a gateway for robust modelling.

- Completing an ethics declaration would enable practitioners to consider their software and algorithms [24, 62, 71, 81].
- The UK Government ought to re-examine the subject benchmark statements for technical courses from secondary to tertiary education and add in best practise and ethics to them [102].

11.4 Practitioners and managers found it difficult to find and retain the right talent

- From the survey responses, it is clear that it is currently difficult to attract the right talent or top talent and that that is a struggle particularly for early stage SMEs undertaking significant development. The Royal Statistical Society [79] is moving towards this with allocation of a designation for Data Scientist, but this is in the development stages currently. The Science Council provides a Chartered Scientist designation, but this is not specifically directed towards modelling practitioners [21]. Therefore, there is a gap that might be filled with a professional body catering to developing and designating modelling practitioners of the future. The link between academia and industry must be a closer collaboration as development moves forward [5, 34, 47, 51, 72, 90, 102, 103].

11.5 Organisational barriers

- Leadership must be involved in these recommendations as no less than a cultural shift is needed to implement them. Leadership training in this area is crucial. Good leadership can create safe spaces for challenge. Within the data analysis, some practitioners report having difficulty knowing which priority to attend to, whether PR or regulatory, this leaves the practitioner potentially addressing the wrong priority or trying to address as many as possible and becoming swamped in competing directions [13, 45, 46, 72, 98].
- Foster an environment of constructive challenge, especially in the public sector. Without constructive challenge of fairness or explainability [39], there is a lack of difficult discussions which are needed to drive this area forward [61, 72, 78].

Acknowledgements The author would like to acknowledge Dr. Ella Haig and Murray McMonies for draft review. In this paper, Dr. Haig was involved in discussions of structure and helped to review early drafts in terms of structure. Murray McMonies has reviewed all drafts for typos and structure and has made some argument critique.

Author contributions The author is fully responsible for the study conception and design. Material preparation and analysis were performed by MO. MO designed the study, obtained ethical approval, and collected and analysed the data. The first and subsequent drafts of the manuscript were written by MO who commented on the previous versions of the manuscript. MO has read and approved the final manuscript. NB—this paper has an abstract that was uploaded as draft over a year ago to PhilPapers. This is not the whole paper or a pre-print; it is just the abstract. However, despite requesting deletion of this abstract, it has not been actioned. As a result, the abstract is still on the pure record of the University of Portsmouth which has been inaccessible for an update for over a year. Despite requesting the deletion of the abstract from this site, it may not have been actioned or actioned yet.

Funding None.

Availability of data and materials It is available in download format from JISC.

Code availability N.A.

Declarations

Conflict of interest None.

Ethical approval Full, available in download from the University of Portsmouth.

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International Conference on Machine Learning. PMLR, pp 60–69 (2018)
2. Appelbaum, D., Kogan, A., Vasarhelyi, M.A.: Big data and analytics in the modern audit engagement: research needs. Audit. J. Pract. Theory **36**(4), 1–27 (2017)
3. Asquith, P.J.: The inevitability and utility of anthropomorphism in description of primate behaviour. Mean. Primate Signals **1984**, 138–176 (1984)
4. Astington, J.W., Baird, J.A.: Why Language Matters for Theory of Mind. Oxford University Press (2005)
5. Axinn, W.G., Pearce, L.D.: Mixed Method Data Collection Strategies. Cambridge University Press (2006)
6. BBC. 2020. Facial recognition use by South Wales Police ruled unlawful. https://tech.newstatesman.com/guestopinion/algorithmic-decision-making (2020)
7. BBC. 2020. Home Office drops 'racist' algorithm from visa decisions. https://www.bbc.co.uk/news/technology-53650758 (2020)

8. Bîgu, D., Cernea, M.-V.: Algorithmic Bias in Current Hiring Practices: An Ethical Examination. In: Proceedings of the International Management Conference, vol. 13. faculty of management, Academy of Economic Studies, Bucharest, Romania, pp 1068–1073 (2019)
9. Binns, R.: Fairness in machine learning: lessons from political philosophy. In: Conference on Fairness, Accountability and Transparency. PMLR, pp 149–159 (2018)
10. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: '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. pp 1–14 (2018)
11. Blacklaws, C.: Algorithms: transparency and accountability. Philos. Trans. Royal Soc. A: Math. Phys. Eng. Sci. **376**(2128), 20170351 (2018)
12. Bosch, N., D'Mello, S.K., Baker, R.S., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., Zhao, W.: Detecting student emotions in computer-enabled classrooms. In: IJCAI. pp. 4125–4129 (2016)
13. Bratasanu, V.: Leadership decision-making processes in the context of data driven tools. Qual.-Access Success **19**, 77–87 (2018)
14. Buhmann, A., Paßmann, J., Fieseler, C.: Managing algorithmic accountability: balancing reputational concerns, engagement strategies, and the potential of rational discourse. J. Bus. Ethics **163**(2), 265–280 (2019)
15. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLR, pp. 77–91 (2018)
16. Cadwalladr, C., Harrison, E.G.: How Cambridge analytica turned Facebook 'likes' into a lucrative political tool, May 2018
17. Chae, Y.: US AI regulation guide: Legislative overview and practical considerations. J. Robot. Artif. Intell. Law **3**(1), 17–40 (2020)
18. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017)
19. Clement-Jones, I.: The government's approach to algorithmic decision-making is broken: here's how to fix it. The Guardian (2020)
20. Coeckelbergh, M.: Language and technology: maps, bridges, and pathways. AI Soc. **32**(2), 175–189 (2017)
21. Science Council. [n.d.]. Chartered Scientist
22. Crane, T.: The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation. Routledge (2015)
23. Crawford, K.: The Atlas of AI. Yale University Press (2021)
24. Crick, J.M., Crick, D.: Angel investors' predictive and control funding criteria: the importance of evolving business models. J. Res. Mark. Entrep. **20**(1), 34–56 (2018)
25. Deeks, A.: The judicial demand for explainable artificial intelligence. Columbia Law Rev. **119**(7), 1829–1850 (2019)
26. Díaz, M., Johnson, I., Lazar, A., Piper, A.M., Gergle, D.: Addressing age-related bias in sentiment analysis. In: Proceedings of the 2018 chi conference on human factors in computing systems. pp 1–14 (2018)
27. Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K.E., Dugan, C.: Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 275–285 (2019)
28. Dunn, P.K., Marshman, M.F.: Teaching mathematical modelling a framework to support teachers' choice of resources. Teach. Math. Appl.: Int. J. IMA **39**(2), 127–144 (2020)
29. Ehsan, U., Riedl, M.O.: Human-centered explainable AI: towards a reflective sociotechnical approach. In: International Conference on Human-Computer Interaction. Springer, pp. 449–466 (2020)
30. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
31. Fast, E., Horvitz, E.: Long-term trends in the public perception of artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31 (2017)
32. Fiori, C., Marzano, V.: Modelling energy consumption of electric freight vehicles in urban pickup/delivery operations: analysis and estimation on a real-world dataset. Transp. Res. Part D: Transp. Environ. **65**, 658–673 (2018)
33. Flores, A.W., Bechtel, K., Lowenkamp, C.T.: False positives, false negatives, and false analyses: a rejoinder to machine bias: there's software used across the country to predict future criminals and it's biased against blacks. Fed. Probat. **80**, 38 (2016)
34. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for datasets. arXiv preprint arXiv:1803.09010 (2018)
35. Gkamas, V., Rigou, M., Paraskevas, M., Zarouchas, T., Perikos, I., Vassiliou, V., Gueorguiev, I., Varbanov, P.: Bridging the skills gap in the data science and internet of things domains: a vocational education and training curriculum (2019)
36. Gogolla, M., Hilken, F., Doan, K.-H.: Achieving model quality through model validation, verification and exploration. Comput. Lang. Syst. Struct. **54**(2018), 474–511 (2018)
37. Goodman, B.W.: A step towards accountable algorithms? Algorithmic discrimination and the European Union general data protection. In: 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona. NIPS Foundation (2016)
38. Grimm, V., Augusiak, J., Focks, A., Frank, B.M., Gabsi, F., Johnston, A.S.A., Liu, C., Martin, B.T., Meli, M., Radchuk, V., et al.: Towards better modelling and decision support: documenting model development, testing, and analysis using TRACE. Ecol. Model. **280**, 129–139 (2014)
39. Hacker, P., Krestel, R., Grundmann, S., Naumann, F.: Explainable AI under contract and tort law: legal incentives and technical challenges. Artif. Intell. Law **2020**, 1–25 (2020)
40. Hamidi, F., Scheuerman, M.K., Branham, S.M.: Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In: Proceedings of the 2018 chi conference on human factors in computing systems, pp 1–13 (2018)
41. Hamon, R., Junklewitz, H., Sanchez, I.: Robustness and explainability of artificial intelligence. Publications Office of the European Union (2020)
42. Hayes, J.C., Kraemer, D.J.M.: Grounded understanding of abstract concepts: the case of STEM learning. Cogn Res: Princ Implic. **2**(1), 1–15 (2017)
43. Hengeveld, G.M., van der Greft-van Rossum J.G.M., de Bie, P.A.F.: Quality assurance models & datasets WENR-WOT: WI0021 Version 1.0. (2021)
44. Holstein, K., Vaughan, J.W., Daumé III, H., Dudik, M., Wallach, H.: Improving fairness in machine learning systems: what do industry practitioners need? In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp 1–16 (2019)
45. Hudson, T.E.: If sages worked in tech: ancient wisdom for future-proof leadership. J. Leadersh. Stud. **13**(4), 43–47 (2020)
46. Johannessen, J.-A.: Knowledge Management for Leadership and Communication: AI, Innovation and the Digital Economy. Emerald Group Publishing (2020)
47. Kallus, N., Zhou, A.: Residual unfairness in fair machine learning from prejudiced data. In: International Conference on Machine Learning. PMLR, pp. 2439–2448 (2018)
48. Kennedy, J.S.: The New Anthropomorphism. Cambridge University Press (1992)

49. Knoppers, B.M., Thorogood, A.M.: Ethics and big data in health. Curr. Opin. Syst. Biol. **4**, 53–57 (2017)
50. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in neural information processing systems, pp. 4066–4076 (2017)
51. Kwak, S.K., Kim, J.H.: Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol **70**(4), 407 (2017)
52. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How we analyzed the COMPAS recidivism algorithm. ProPublica (5 2016) **9**(1), 3 (2016)
53. Lawless, W.F., Mittu, R., Sofge, D., Hiatt, L.: Artificial intelligence, autonomy, and human-machine teams: interdependence, context, and explainable AI. AI Mag **40**(3), 5–13 (2019)
54. Lee, M.K.: Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc **5**(1), 2053951718756684 (2018)
55. Lenk, H.: Ethics of responsibilities distributions in a technological culture. AI Soc **32**(2), 219–231 (2017)
56. Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al.: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit. Health **1**(6), e271–e297 (2019)
57. Louart, C., Couillet, R.: A concentration of measure and random matrix approach to large dimensional robust statistics. arXiv preprint arXiv:2006.09728 (2020)
58. Lum, K., Isaac, W.: To predict and serve? Significance **13**(5), 14–19 (2016)
59. Malik, V., Singh, S.: Tools, strategies & models for incorporating software quality assurance in risk oriented testing. Orient. J. Chem. **10**(3), 603–611 (2017)
60. Martin, K.: Ethical implications and accountability of algorithms. J. Bus. Ethics **160**(4), 835–850 (2019)
61. Mashelkar, R.A.: Impact of science, technology and innovation on the economic and political power. AI Soc. **32**(2), 243–251 (2017)
62. Metcalf, J., Moss, E., Watkins, E.A., Singh, R., Elish, M.C.: Algorithmic impact assessments and accountability: the co-construction of impacts. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 735–746 (2021)
63. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**(4), 2141–2168 (2020)
64. Muralidhar, N., Islam, M.R., Marwah, M., Karpatne, A., Ramakrishnan, N.: Incorporating prior domain knowledge into deep neural networks. In: 2018 IEEE international conference on big data (big data). IEEE, pp. 36–45 (2018)
65. Nagel, T.: What is it like to be a bat? Philos. Rev. **83**(4), 435–450 (1974)
66. Narayanan, A.: Translation tutorial: 21 fairness definitions and their politics. In Proc. Conf. Fairness Accountability Transp., vol. 1170. New York, USA (2018)
67. Neri, H., Cozman, F.: The role of experts in the public perception of risk of artificial intelligence. AI Soc. **2019**, 1–11 (2019)
68. Nicoll, P.: Audit in a Democracy: the Australian Model of Public Sector Audit and Its Application to Emerging Markets. Routledge (2016)
69. Information Commissioner's Office and the Turing Institute. Explaining decisions made with AI (2020)
70. Oldfield, M.: AI: anthropomorphism and dehumanisation. In: 5th Digital Geographies Research Group Annual Symposium 2021: Where Next for Digital Geographies? Pathways and Prospects (2021)
71. Oldfield, M., Gardner, A., Smith, A.L., Steventon, A., Coughlan, E.: Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00069-w
72. Oldfield, M., Haig, E.: Analytical modelling and UK Government policy. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00078-9
73. Peters, J.: IBM will no longer offer, develop, or research facial recognition technology. The Verge, June 8 (2020)
74. Peters, R.G., Covello, V.T., McCallum, D.B.: The determinants of trust and credibility in environmental risk communication: an empirical study. Risk Anal **17**(1), 43–54 (1997)
75. QAA.: Subject Benchmark Statement - Computing. https://www.qaa.ac.uk/docs/qaa/subject-benchmarkstatements/subject-benchmark-statement-computing.pdf?sfvrsn=ef2c881_10 (2019)
76. Rader, E., Gray, R.: Understanding user beliefs about algorithmic curation in the Facebook news feed. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 173–182 (2015)
77. Reddy, E., Cakici, B., Ballestero, A.: Beyond mystery: putting algorithmic accountability in context. Big Data Soc **6**(1), 205 (2019)
78. Robinson, A., Glover, P.: Developments in the quality assurance of government models used to support business critical decisions. In: Proceedings of the Operational Research Society Simulation Workshop. pp. 176–181 (2014)
79. RSS. [n.d.]. Data Scientist
80. Sabillon, R., Serra-Ruiz, J., Cavaller, V., Cano, J.: A comprehensive cybersecurity audit model to improve cybersecurity assurance: the cybersecurity audit model (CSAM). In: 2017 International Conference on Information Systems and Computer Science (INCISCOS). IEEE, pp. 253–259 (2017)
81. Safdar, N.M., Banja, J.D., Meltzer, C.C.: Ethical considerations in artificial intelligence. Eur. J. Radiol. **122**, 108768 (2020)
82. Samuel, G., Diedericks, H., Derrick, G.: Population health AI researchers' perceptions of the public portrayal of AI: a pilot study. Public Underst. Sci. **30**(2), 196–211 (2021)
83. Schubert, A., Ahsbahs, C.: The ESCB quality framework for European statistics. Austrian J. Stat. **44**(2), 3–11 (2015)
84. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59–68 (2019)
85. Shen, J., Zhang, C.J.P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S.Y., Fang, P.-H., Ming, W.-K.: Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR Med Inform **7**(3), e10010 (2019)
86. Shin, D.: User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explainability. J. Broadcast. Electron. Media **64**(4), 541–565 (2020)
87. Shin, D., Park, Y.J.: Role of fairness, accountability, and transparency in algorithmic affordance. Comput. Hum. Behav. **98**, 277–284 (2019)
88. Sokol, K., Flach, P.: Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 56–67 (2020)
89. The Times.: Police scrap artificial intelligence tool to predict violence. https://www.thetimes.co.uk/article/policescrap-artificial-intelligence-tool-to-predict-violence-zdln8bgz0 (2020)
90. Toyama, K.: From needs to aspirations in information technology for development. Inf Technol Dev **24**(1), 15–36 (2018)

91. HM Treasury. Review of quality assurance of government analytical models. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/206946/review_of_qa_of_govt_analytical_models_final_report_040313.pdf (2013)
92. HM Treasury, H.M.: The Aqua Book: guidance on producing quality analysis for government. https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-qualityanalysis-for-government (2015)
93. Turiel, J.D., Aste, T.: Peer-to-peer loan acceptance and default prediction with artificial intelligence. Royal Soc. Open Sci. **7**(6), 191649 (2020)
94. Veale, M., Van Kleek, M., Binns, R.: Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In: Proceedings of the 2018 chi conference on human factors in computing systems, pp. 1–14 (2018)
95. Wood, S.: Review of quality assurance of government analytical models. https://www.smh.com.au/national/a-lot-of-people-are-sleepwalking-into-it-the-expert-raising-concerns-over-ai-20210714-p589qh.html (2021)
96. Woodruff, A., Fox, S.E., Rousso-Schindler, S., Warshaw, J.: A qualitative exploration of perceptions of algorithmic fairness. In: Proceedings of the 2018 chi conference on human factors in computing systems. pp. 1–14 (2018)
97. Xanthopoulos, P., Pardalos, P.M., Trafalis, T.B.: Robust Data Mining. Springer, New York (2012)
98. Yammarino, F.J., Salas, E., Serban, A., Shirreffs, K., Shuffler, M.L.: Collectivistic leadership approaches: putting the "we" in leadership science and practice. Ind. Organ. Psychol. **5**(4), 382–402 (2012)
99. Yao, B., Vasiljevic, M., Weick, M., Sereno, M.E., O'Donnell, P.J., Sereno, S.C.: Semantic size of abstract concepts: It gets emotional when you can't see it. PLoS ONE **8**(9), e75000 (2013)
100. Zhai, Y., Yan, J., Zhang, H., Lu, W.: Tracing the evolution of AI: conceptualization of artificial intelligence in mass media discourse. Inf. Discov. Deliv. **48**(3), 137–149 (2020)
101. Zhao, X., Phillips, E.K., Malle, B.F.: Beyond anthropomorphism: differentiated inferences about robot mind from appearance. ACR North American Advances (2019)
102. Oldfield, M.: Towards pedagogy supporting ethics in modelling. J. Humanist. Math. **12**(2), 128–159 (2022)
103. Oldfield, M., McMonies, M., Haig, E.: The future of condition based monitoring: risks of operator removal on complex platforms. AI Soc. (2022). https://doi.org/10.1007/s00146-022-01521-z