Fairness in AI and Its Long-Term Implications on Society

Fairness in AI and Its Long-Term Implications on Society

Ondrej Bohdal[1,*], **Timothy Hospedales**[1,2], **Philip H.S. Torr**[3], **Fazl Barez**[1,3,*]
[1] School of Informatics, University of Edinburgh [2] Samsung AI Center, Cambridge
[3] Department of Engineering Science, University of Oxford [*] Main contributors

Abstract

Successful deployment of artificial intelligence (AI) in various settings has led to numerous positive outcomes for individuals and society. However, AI systems have also been shown to harm parts of the population due to biased predictions. We take a closer look at AI fairness and analyse how lack of AI fairness can lead to deepening of biases over time and act as a social stressor. If the issues persist, it could have undesirable long-term implications on society, reinforced by interactions with other risks. We examine current strategies for improving AI fairness, assess their limitations in terms of real-world deployment, and explore potential paths forward to ensure we reap AI's benefits without harming significant parts of the society.

1 Introduction

AI approaches offer excellent performance in many practically important problems [1–3], but they can give biased and unfair predictions [4–6]. AI is increasingly often deployed to high-stakes applications [7–9], where unfair predictions can lead to substantial disadvantage or harm to parts of the population. For example, AI has been used for deciding who to select for interviews [10, 11], who should be given a mortgage [12] or who is more likely to repeat crime after leaving prison [7, 13]. Unfair decisions in such key areas can have a significant impact on one's future.

We study the long-term social implications of unfair AI, from the perspective of continuous bias amplification stemming from new AI models trained on increasingly biased data. Biased AI models lead to biased outcomes in the real-world, which will serve as biased data for training new, more biased, AI models. Additionally parts of the population can experience bias from several sources, e.g. hiring and healthcare, and these combined can also put certain groups in increasingly large disadvantage over time. Overall this may lead to a feedback loop where new AI models become more and more biased.

If parts of the population are systematically marginalized because of biased AI models, they can be under severe stress, and they may try to resolve the situation by protesting against the deployment of such AI systems. If the institutions find it challenging to stop using biased AI, e.g. due to lack of employees or resources more broadly, disadvantaged groups may resort to escalating the situation. In this sense lack of AI fairness can act as a social stressor if the issues are prevalent and not addressed.

Deployment of insufficiently fair AI systems would likely be only one of several social stressors. Consequently, we study the interaction with other stressors, especially climate change that has increasingly significant impact on the society. The interaction among multiple social stressors can reinforce each other and result in more extensive tension in the society.

In addition to studying the social implications, we also investigate what approaches are being developed to improve AI fairness. We give particular focus on real-world deployment of fair AI models and identify that lack of fairness generalization across data distribution shifts can be a

3rd Annual Stanford Existential Risks Conference, 2023

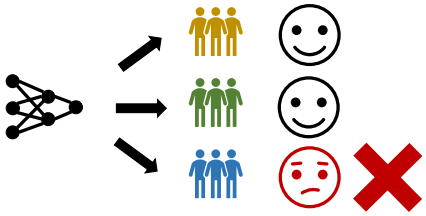

Figure 1: We focus on the topic of fairness where we want to ensure that all groups receive unbiased and equal treatment so that no groups are harmed because of using AI.

key challenge. We discuss approaches for robust fairness, but we also discuss approaches from areas related to out-of-distribution robustness, including domain generalization and adaptation. We conclude by giving recommendations for further research that can be taken to improve the real-world impact of fairness.

2 Definitions of Fairness

Researchers have proposed a variety of ways to define fairness [4], and some of the most common include equalized odds [5], equal opportunity [5] and demographic parity [6]. The unifying theme of these metrics is that we want to ensure the same or similar probability of selected outcomes across all of the considered groups, which we illustrate in Figure 1.

Equalized odds [5] for predictor $\widehat{Y}$, target Y and protected attribute A are defined for the binary case as:

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=y\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=y\right\}, \quad y \in\{0,1\}.$$

The definition means $\widehat{Y}$ has equal true positive rate for demographics $A=0$ and $A=1$ if the outcome is $y=1$, and equal false positive rates if the outcome is $y=0$.

Equal opportunity [5] is a relaxed alternative of equalized odds as it only requires non-discrimination within the advantageous outcome group:

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=1\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=1\right\}.$$

Compared to earlier demographic parity metric [6], the benefit of equalized odds and equal opportunity is that they do not require independence from the protected attribute [5]. More broadly when deciding which metric to use, it is key to consider the suitability for the specific application [14].

In addition to specialized fairness metrics, we can monitor the worst-case performance alongside the average performance [15, 16]. More specifically we can measure the performance on the most challenging group [15] or if the notion is less clear, we can use e.g. the most challenging 10% of the examples used for evaluation [16]. Such way of evaluation can also be used for settings where we want to ensure fairness when deploying AI systems across different scenarios. It is related to Max-Min fairness [17] where a model with smaller worst-case error is seen as fairer.

We further consider a stronger notion of fairness that is important when deploying models to the real-world: fairness that is robust under various data distribution shifts. AI models should not discriminate against any of the subgroups when deployed to real-world "in-the-wild" scenarios. We illustrate robust fairness in Figure 2, and we will also consider this notion when discussing current solutions towards fairness.

3 Social Implications of AI Fairness

We begin our analysis of social implications of AI fairness by introducing several high-stakes real-world examples where AI has been used already. We will then present a self-reinforcing feedback

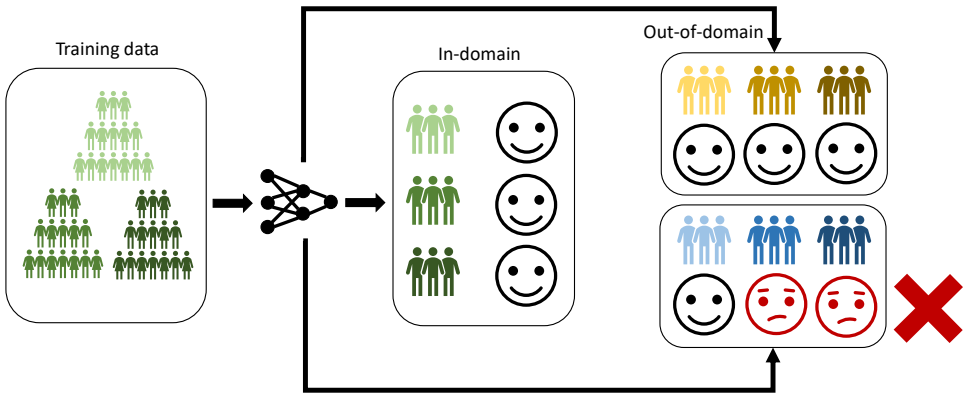

Figure 2: When deploying AI models to the real world, it is crucial to ensure the models are robust and generalize fairness also to out-of-domain situations.

loop mechanism where biased AI systems lead to biased outcomes, which then act as input data for further training of new AI systems. Over long periods of time this may lead to increasingly systemic social and economic marginalization of parts of the population. Such systemic marginalization could later become a substantial social stressor.

3.1 High-Stakes Real-World Applications

With the possibility to automate various time-consuming tasks and potentially improve upon imperfections of human decision-making, AI has been used for a number of high-stakes applications where fairness is important [4]. However, in many cases it has already been identified that the AI is unfair and causes harm to certain groups. High-stakes real-world applications where fairness matters and has already been compromised include the following:

- **Hiring for jobs**: biased AI has been used in the context of hiring in multiple ways, including filtering of CVs [10, 11], evaluating video interviews [18] and delivering advertisements promoting jobs [19]. Type of employment has a large impact on one's future, so it is key to ensure certain groups are not systematically disadvantaged (or given an advantage).
- **Finance**: AI can simplify the task of assessing if someone is likely to repay a loan or a mortgage, so such systems have already been deployed in practice. It has been shown that systems for making decisions about loans [20] or mortgages [12] can be significantly biased, for example making applicants of colour 40 to 80% more likely to be denied mortgage application compared to white applicants [12]. If a group of certain characteristics is unable to get a mortgage and is forced to rent, it can have a large impact on their well-being, especially if it means they have to find new accommodation often.
- **Public safety**: unfair AI systems have been used in various public safety contexts, including sentencing decisions [13, 7] and children welfare [21]. More specifically, AI has been used to predict the risk of recidivism as part of the COMPAS system [13], and also as part of a tool applied to the particularly sensitive case of predicting juvenile recidivism [7]. Biased AI has also been used in the context of children welfare to perform screening of referrals for child protection [21].
- **Healthcare**: biased AI systems have been deployed for multiple healthcare applications. For example, health-management systems [22] have been shown to assign the same risk to black patients that are sicker than white patients. Biased AI has also led to underdiagnosis of under-served patient populations when applying AI to chest radiographs [9], and it also resulted in gender-biased computer-aided diagnosis [8].

Additionally, it is not only the high-stakes situations where AI has the potential to discriminate and treat people unfairly. There are also many situations where unfair AI can cause inconvenience. However, these can potentially act as a reminder that the person may have been treated unfairly by AI in some of the high-stakes situations.

Face recognition is one of the key areas that exemplifies such scenarios and shows the need for fair and robust AI models. For instance, earlier facial processing systems from leading tech companies performed significantly worse on black women than on white men [23, 24]. Widespread use of such models could lead to frequent inconveniences, for example if face recognition technology were utilized for workplace access. Other adverse examples include:

- Identifying criminals in public spaces using facial recognition technology, where the risk of misidentification could lead to reputational damage or legal charges for the affected individual.
- Implementing face recognition in airports, where failure to recognize an individual could result in additional time spent on alternative verification methods, such as waiting in a separate queue.

The concept of fairness is also important in the context of generative language models such as GPT-4 [25], LaMDA [26] and LLaMA [27], because the generated content can influence people and have real-world impact. We want to ensure the generated content is not biased and does not include prejudices about any parts of the population. Further we want to ensure that safeguards in the models are robust across different languages and cultures so that we do not risk, for example, harming parts of the society in countries that speak lower-resource languages.

3.2 Self-Reinforcing Feedback Loop

Deployed AI models influence the society, and the outcomes form part of the training data for a new generation of AI models. If the initial models are biased, they produce biased outputs that will be used for training newer models. It has been shown that AI models can amplify biases [28, 29], which means new AI models would be biased even more due to training on increasingly biased data. We illustrate this in Figure 3, where we show the resulting feedback loop. Because many biases become evident only after wide deployment of the system, it is key to consider what long-term implications AI-amplified biases could have.

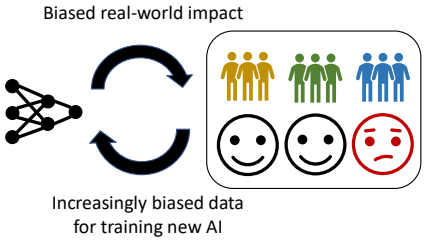

Figure 3: Biased real-world outcomes lead to increasingly biased data for training new AI models, resulting in a self-enforcing feedback loop.

Let us explain the bias amplification on examples. For example, if jobs hiring decision AI is based on past hiring decisions, then bias against subgroups in the past can reinforce to more bias in the future. If criminal sentencing AI is biased against a subgroup, that subgroup has longer prison sentences, which may make them harder to re-integrate with society after release. This may increase their likelihood of repeat crime recidivism, which will be data that increases the bias of the sentencing-decision AI the next time it is trained.

A whole ecosystem of AI models that happened to be biased against a particular subgroup could lead to persistently worse social and economic outcomes for that subgroup. More specifically, it could lead to worse jobs, worse access to finance, longer sentences for equivalent crimes, worse health due to worse medical treatment, worse educational outcomes if the education AI is biased. These

biases then reinforce each other as e.g. worse health reinforces worse education and jobs. Over long periods of time that subgroup could become increasingly systemically socially and economically marginalized, which could become a significant social stressor.

There is a risk of compounding of negative effects due to the feedback potential between the AI system decisions and real-world data, which affects the training data for the next round of AI training. The level of bias may become increasingly more difficult to tolerate. Moreover, as the technology becomes widespread the infrastructure will be built around it, so it will be challenging to remove it from use even as people protest against it. This would ultimately create tensions.

3.3 Fairness Risk

We present a toy model to estimate the fairness risk at time t after deployment of the first AI systems. The formula models the compounding of biases over time (similar to compounding of interest rates):

$$\text{risk} \sim \beta^t,$$

where β is the bias amplification rate of the AI models. When no bias amplification happens, $\beta = 1$, but because AI models have been shown to increase biases, typically $\beta > 1$. Over time the biases would amplify each other via repeated training of new AI systems on increasingly biased data. Such amplification of biases can lead to dissatisfaction with the AI systems and can act as a social stressor. Ideally we would like to have $\beta \leq 1$ so that AI decreases biases in the society.

3.4 Challenges with Avoiding AI Automation

Government budgets are typically tight [30], and ways to save resources are sought. AI offers ways to decrease costs [31–33] and if there is a crisis, institutions may be more likely to use experimental approaches that have not been fully assessed. If the technology is likely to benefit most of the population, it may be difficult to argue against its deployment, especially if it is hard to measure potential harm it can cause. In these cases, it is important to recognize effects across different groups. One can try to mitigate them by investing more resources into making the technology fair and robust. It is also important to note that AI-based solutions may become deeply embedded in the government software infrastructure and removing them once new significant biases are identified or if people start protesting, can be challenging.

In addition, automation using AI is not only about savings, it may also be inevitable due to shortages of employees to perform specific jobs [34]. For example, shortages have been reported in areas as diverse as construction, social work and transportation [35]. Shortages may not always be resolved by increasing the budgets because some jobs require hard-to-obtain skills or cause significant amount of distress, so using AI would be desirable also because of non-monetary reasons.

3.5 Interaction with Other Risks and Long-Term Implications

Unfair AI would be only one factor that would contribute to tension in the society. Another key driver of tension is likely to be climate change [36] and its implications [37], which include rising food prices or having to cover the costs of disaster responses. For example, large-scale drought in Syria is thought to have contributed to social stressors, which eventually led to an uprising in 2011 [38]. Syria is in civil war since 2011, already for more than a decade [39, 40]. The case of Syria shows that a combination of multiple stressors that lead to uprisings can result in a long-term civil war, which is a prime example of country's crisis. More broadly systemic unfairness, inequality and marginalization of parts of the population has a record of leading to radicalization [41], uprisings and in some cases destabilization of societies. High-profile examples where these were likely to be a factor include the French revolution [42], Indian independence movement [43] and recently the Arab Spring [44].

A combination of multiple social stressors such as climate change and biased AI are likely to reinforce each other, which we illustrate in Figure 4. Persistent deployment of biased AI in high-stakes applications can lead to increasing levels of tension in the society and erode trust in the institutions. At a certain level it may be significant enough that in interaction with other social stressors such as climate change it escalates. It is crucial to try to mitigate the social stressors to avoid any compounding effects.

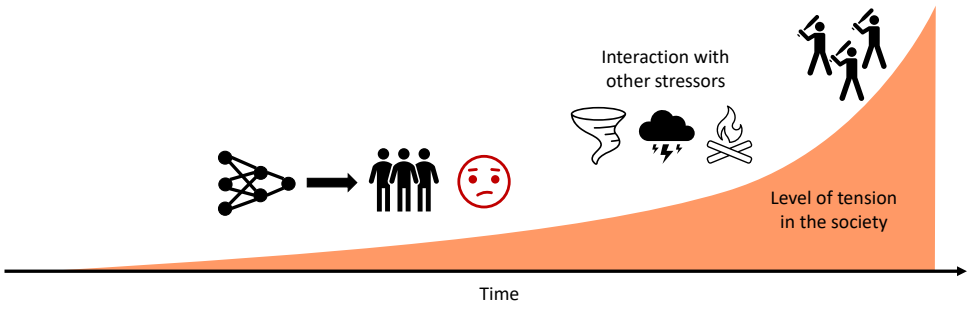

Figure 4: Biased AI outcomes in interaction with other social stressors can lead to increasing levels of tension in the society and escalate if not mitigated.

Biased AI, climate change and other social stressors have the potential to be commonplace across many nations, which increases the importance of developing strategies that can mitigate them. Potential negative implications of social stressors can be alleviated, for example, by giving significant focus and funding to research that can develop suitable solutions.

4 How to Improve AI Fairness?

4.1 Approaches for Fairness

A large number of approaches for fairness has been proposed [4, 14], reflecting the significant importance of the field. Many of the recent methods try to alleviate bias as part of training the models, also known as in-processing [4, 14]. Key in-processing families of bias mitigation methods include:

- **Subgroup rebalancing** [45, 46] that over-samples minority groups and down-samples majority groups,
- **Domain independence** [47, 48] that uses separate classifiers for different subgroups,
- **Adversarial training** [49–51] that tries to train representations that make it difficult to identify different groups,
- **Disentanglement** [52, 53] methods that separate sensitive attributes and the useful attributes when constructing the representations.

Other families of methods for improving fairness also exist. [14] have identified that domain generalization approaches [54–56] can be useful for improving fairness. Domain generalization methods try to learn representations that directly generalize to new out-of-domain situations without any adaptation, which relates to the goal of obtaining strong performance across different groups present in the population. Further, pre-processing methods [57] try to remove bias from the dataset before training, for example by distorting the data [57] as simply removing the sensitive attributes has been shown to be insufficient [4]. Post-processing methods [58] modify the predictions of an already trained model to improve fairness with respect to the sensitive attributes.

4.2 Fairness Under Data Distribution Shifts

When deploying AI models to the real-world, it is key to ensure that the key properties of AI models hold in the presence of real "in-the-wild" data. Such data are likely to come also from data distributions different from ones seen during training, so robustness against distribution shifts is crucial. For example, healthcare AI systems can be trained on data from selected prestigious US hospitals, and deployed to hospitals of various quality across the US.

However, it has been shown that most existing fairness methods are only designed for in-domain settings and fail when data distribution changes [59, 60, 14]. Several approaches have been developed to tackle the challenge of fairness under distribution shift [61, 62, 59], but these consider adaptation to a specific domain, with only [63] presenting an approach that generalizes across domains. If

institutions only consider if an AI system is fair for in-domain data, such AI models may still lead to significant biases when deployed in the real-world and cause harm to parts of the population.

It has been identified [14] that domain generalization approaches [54–56] can offer competitive performance in terms of fairness, so improving fairness of domain generalization methods can be a good way forward. However, it has also been shown that domain generalization is a challenging problem on its own [64], with many approaches performing similarly as simple training across many domains [65] if following a fair evaluation protocol. As a result, domain adaptation approaches that adapt pre-trained models to local data distributions can be more successful in terms of maintaining fairness and strong performance. Source-free domain adaptation [66–69] in particular can be practically valuable as it adapts a pre-trained model solely using unlabelled target domain data, without access to the source data. Efficient feed-forward approaches that perform adaptation without back-propagation [16, 70–72] can be especially useful on deployed devices.

4.3 Evaluation

Various benchmarks have been developed or repurposed for evaluating fairness. Key tabular fairness datasets include COMPAS [73], Adult Census [74, 75] and Diabetes [76], some of which are also available within the popular Fairlearn library [77]. Common computer vision fairness datasets include CIFAR-10S [47], CelebA [78] and IMDB face dataset [79]. Medical imaging datasets [80–82] have also been used extensively for evaluating fairness, with MEDFAIR [14] providing a suite of benchmarks to provide rigorous evaluation of fairness algorithms, including in-domain and out-of-domain scenarios.

Long term we believe it is key to develop new more extensive benchmarks that test both in-domain and out-of-domain scenarios, similar in scope to MEDFAIR [14] but covering various areas for which AI fairness is crucial. Because it has been observed that real-world datasets are often biased, synthetic datasets may be highly useful in the future. Synthetic data would enable us to design what unbiased outcomes look like and train models on them, improving fairness and robustness [83]. Once the model is trained with synthetic data, it can be fine-tuned using curated real-world data that does not need to be as plentiful.

4.4 Recommendations for Research and Deployment

Considering the significant impact that AI fairness can have on the future of our society, there are several steps that both AI researchers and practitioners can follow to mitigate negative impact:

- Develop a deeper understanding of the distinct role of different sources of algorithmic bias (such as differing sub-population size and label frequencies, label-noise imbalance, spurious correlations [4, 14]), and how these interact with various notions of fairness relevant in different social contexts (equality of opportunity, max-min fairness).
- Develop a science of iterative bias amplification that will help us understand how decisions made by current AI systems (which determine the training set of future AI systems) affect the evolution of AI bias and fairness in the long run.
- Develop new benchmarks and simulators that will allow us to more rigorously benchmark AI for bias and fairness, both in-domain and out-of-domain, and for single and multiple rounds of training.
- Develop new foundational synthetic datasets that can be used for fair pre-training of AI models.
- Develop new AI training and inference algorithms that lead to robust improvements in fairness. These should address both single and multiple rounds of training, as well as ensuring fairness when deployed to real-world situations with distribution shift compared to the training data.
- Monitor deployed AI systems as they influence people in the real-world to ensure any significant biases are identified and resolved early.

More speculatively AI could also be used to mitigate biases present in the society, for example by evaluating content for bias and alerting users about it.

5 Discussion and Broader Recommendations

5.1 Importance of Fair AI systems

In order to create a fair and inclusive society, AI developers and researchers need to focus on enhancing the fairness and robustness of their models. This encompasses several key aspects:

- Regularly testing and refining AI models to minimize biases and optimize performance across various populations.
- Promoting transparency in AI development by allowing external audits and assessments of model performance.
- Engaging with stakeholders, such as affected communities, to better understand potential risks and address any concerns.
- Implementing clear guidelines and regulations to direct the ethical use of AI technologies, particularly in high-stakes applications such as healthcare, finance and law enforcement.
- Investing in research and tools for evaluating the capabilities of existing systems to prevent undesirable behaviour. This may involve "red teaming" or other role-playing techniques that can help identify potential unintended consequences of current methods, which could become real issues in the near future if not addressed.

5.2 Role of Governments and Organizations

Governments and organizations are pivotal in promoting AI fairness and addressing its potential societal consequences. Their involvement can manifest in various ways, including:

- **Policy development:** establishing and enforcing guidelines that ensure AI systems are designed and deployed responsibly. For example, creating regulations that mandate transparency in AI decision-making processes or setting data privacy and security standards. Specific interventions could include requiring human-in-the-loop for critical applications and more broadly ensuring that institutions only procure systems that have been extensively evaluated in terms of bias.
- **Research support:** allocating funding and resources for research into AI fairness, robustness, ethics, and inclusivity. This may include establishing research centres or providing grants to support projects focused on AI safety and fairness. For example, funding could be provided for initiatives such as development of foundational fair datasets that people can use for training models that will be deployed in real-world applications.
- **Public awareness:** initiate public awareness campaigns to educate citizens about AI technologies' potential risks and benefits, and their impact on society. Such campaigns may include, but are not limited to, educational seminars and social media campaigns that aim to inform the public about AI advancements and their implications, especially unintended consequences.
- **Collaborative efforts:** facilitating collaboration between AI developers, researchers, and affected communities to ensure diverse perspectives are represented in AI development. This can be achieved through creating platforms for dialogue, hosting conferences or workshops, and encouraging partnerships between various stakeholders in the AI ecosystem.

6 Conclusion

In this paper we have investigated the long-term implications of unfair AI systems. We have identified that a feedback loop that leads to increasingly large biases can arise as biased AI models impact the population and new AI models are trained on such outcomes. Over longer time horizons, increasing levels of systemic unfairness can act as a social stressor and trigger protests. We have discussed real-world limitations of existing AI systems designed to be fair and suggested steps that can be taken to improve the situation. Overall we believe that thanks to the significant interest from both the ML community and institutions deploying AI systems, potential severe risks stemming from biased AI systems can be avoided, but carefulness and extensive further research will be key.

Acknowledgements

We are greatful to Charlotte Siegmann, Shahar Avin and Atoosa Kasirzadeh for their valuable feedback on our earlier draft. Their insights and comments have greatly improved the quality of this work.

References

[1] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare. In *KDD*, 2015.

[2] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, 2012.

[3] Javier Andreu-Perez, Fani Deligianni, Daniele Ravi, and Guang-Zhong Yang. Artificial intelligence and robotics. Technical report, UK-RAS Network, 2018.

[4] Ninareh Mehrabi, Fred Morstatter, Nripsuta Ani Saxena, Kristina Lerman, and A. G. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 2019.

[5] Moritz Hardt, , Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.

[6] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS*, 2012.

[7] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in Catalonia. In *International Conference on Artificial Intelligence and Law*, 2019.

[8] Agostina J. Larrazabal, Nicolas Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 2020.

[9] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B A McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 2021.

[10] Lee Cohen, Zachary C. Lipton, and Yishay Mansour. Efficient candidate screening under multiple tests and implications for fairness. In *Symposium on Foundations of Responsible Computing*, 2020.

[11] Miranda Bogen and Aaron Rieke. Help wanted: an examination of hiring algorithms, equity, and bias. 2018.

[12] Emmanuel Martinez and Lauren Kirchner. The secret bias hidden in mortgage-approval algorithms, 2021.

[13] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 2018.

[14] Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking fairness for medical imaging. In *ICLR*, 2023.

[15] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: learning to adapt to domain shift. In *NeurIPS*, 2021.

[16] Ondrej Bohdal, Da Li, Shell Xu Hu, and Timothy Hospedales. Feed-forward source-free latent domain adaptation via cross-attention. In *ICML Pre-training Workshop*, 2022.

[17] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning. In *NIPS*, 2020.

[18] Aislinn Kelly-Lyth. Challenging biased hiring algorithms. *Oxford Journal of Legal Studies*, 2021.

[19] Anja Lambrecht and Catherine Tucker. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 2019.

[20] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P. Mathur. Multi-objective evolutionary algorithms for the risk-return trade-off in bank loan management. *International Transactions in Operational Research*, 2002.

[21] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *FAccT*, 2018.

[22] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 2019.

[23] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*, 2018.

[24] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *AIES*, 2020.

[25] OpenAI. GPT-4 technical report. In *arXiv*, 2023.

[26] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. LaMDA: Language models for dialog applications. In *arXiv*, 2022.

[27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. In *arXiv*, 2023.

[28] Kirsten Lloyd. Bias amplification in artificial intelligence systems. In *arXiv*, 2018.

[29] Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron B. Adcock. A systematic study of bias amplification. In *ArXiv*, 2022.

[30] Mika Tujula and Guido Wolswijk. What determines fiscal balances? an empirical investigation in determinants of changes in oecd budget balances. *SSRN Electronic Journal*, 2004.

[31] George Atalla and Mark MacDonald. How AI can help governments manage their money better. Technical report, 2019.

[32] Peter Viechnicki and William D. Eggers. How much time and money can AI save government? Technical report, 2017.

[33] Niklas Berglind, Ankit Fadia, and Tom Isherwood. The potential value of AI—and how governments could look to capture it. Technical report, 2022.

[34] Martin Ford. Robots: stealing our jobs or solving labour shortages?, 2021.

[35] Brigid Francis-Devine and Isabel Buchanan. Skills and labour shortages. Technical report, House of Commons Library, 2023.

[36] Institute for Economics & Peace. Ecological threat register. Technical report, 2020.

[37] Peter F. Nardulli, Buddy Peyton, and Joseph Bajjalieh. Climate change and civil unrest: The impact of rapid-onset disasters. *The Journal of Conflict Resolution*, 2015.

[38] Colin P. Kelley, Shahrzad Mohtadi, Mark A. Cane, Richard Seager, and Yochanan Kushnir. Climate change in the fertile crescent and implications of the recent Syrian drought. *Proceedings of the National Academy of Sciences*, 2015.

[39] Philip Loft, Georgina Sturge, and Esme Kirk-Wade. The Syrian civil war: Timeline and statistics . In *Commons Library Research Briefing*, 2022.

[40] Hilly Moodrick-Even Khen, Nir T. Boms, and Sareta Ashraph. *Introduction: An overview of stakeholders and interests*, pages 1–8. Cambridge University Press, 2020.

[41] Kees van den Bos. Unfairness and radicalization. *Annual Review of Psychology*, 2020.

[42] Alexis Tocqueville. *The old regime and the revolution*. Harper and Brothers, New York, USA, 1856.

[43] Bipan Chandra. *India's struggle for independence, 1857-1947*. Penguin Books, New Delhi, India, 1989.

[44] Mark L. Haas. *The Arab Spring: The hope and reality of the uprisings*. Routledge, New York, USA, 2017.

[45] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002.

[46] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, 2022.

[47] Zeyu Wang, Klint Qinami, Ioannis Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020.

[48] Amelie Royer and Christoph H. Lampert. Classifier adaptation at prediction time. In *CVPR*, 2015.

[49] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2018.

[50] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. In *ICLR*, 2019.

[51] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *CVPR*, 2019.

[52] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *CVPR*, 2021.

[53] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *ECCV*, 2020.

[54] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020.

[55] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *NeurIPS*, 2021.

[56] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021.

[57] Sajad Khodadadian, AmirEmad Ghassami, and Negar Kiyavash. Impact of data processing on fairness in supervised learning. In *International Symposium on Information Theory (ISIT)*, 2021.

[58] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *NIPS*, 2017.

[59] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H. Chi. Transfer of machine learning fairness across domains. In *NeurIPS AI for Social Good Workshop*, 2019.

[60] Alan Mishler and Niccolò Dalmasso. Fair when trained, unfair when deployed: Observable fairness measures are unstable in performative prediction settings. In *NeurIPS Algorithmic Fairness through the Lens of Causality and Privacy Workshop*, 2022.

[61] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *FAccT*, 2021.

[62] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. Robust fairness under covariate shift. In *AAAI*, 2020.

[63] Thai-Hoang Pham, Xue Zhang, and Ping Zhang. Fairness and accuracy under domain generalization. In *ICLR*, 2023.

[64] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

[65] Vladimir Vapnik. *Statistical learning theory*. 1998.

[66] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.

[67] Jogendra Nath Kundu, Naveen Venkat, Rahul M, and R. Venkatesh Babu. Universal source-free domain adaptation. In *CVPR*, 2020.

[68] Masato Ishii and Masashi Sugiyama. Source-free domain adaptation via distributional alignment by matching batch normalization statistics. In *arXiv*, 2021.

[69] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *NeurIPS*, 2021.

[70] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *NeurIPS*, 2020.

[71] Ondrej Bohdal, Da Li, and Timothy Hospedales. Feed-forward source-free domain adaptation via class prototypes. In *ECCV OOD-CV Workshop*, 2022.

[72] Ondrej Bohdal, Da Li, and Timothy Hospedales. Label calibration for semantic segmentation under domain shift. In *ICLR Workshop on Trustworthy ML*, 2023.

[73] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity, 2016.

[74] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *NeurIPS*, 2021.

[75] Ronny Kohavi and Barry Becker. Adult data set, 1996.

[76] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan Luis Olmo, Sebastián Ventura, Krzysztof J. Cios, and John N. Clore. Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014.

[77] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical report, Microsoft, 2020.

[78] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[79] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *ICCV Workshops*, 2015.

[80] Jeremy A. Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David Andrew Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, C. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and A. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019.

[81] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *Scientific Data*, 2019.

[82] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *CVPR*, 2021.

[83] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data - what, why and how? In *ArXiv*, 2022.