Trustworthy AI in the Age of Pervasive Computing and Big Data

Kumar, Abhishek

IEEE
2020

Kumar , A , Braud , T , Tarkoma , S & Hui , P 2020 , Trustworthy AI in the Age of Pervasive Computing and Big Data . in IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops) . IEEE , 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops) , Austin, TX , United States , 23/03/2020 . https://doi.org/10.1109/PERCOMWORKSHOPS48775.2020.9

http://hdl.handle.net/10138/318607
https://doi.org/10.1109/PERCOMWORKSHOPS48775.2020.9156127

acceptedVersion

Trustworthy AI in the Age of Pervasive Computing and Big Data

Abhishek Kumar*, Tristan Braud†, Sasu Tarkoma*, Pan Hui*†
*Department of Computer Science, University of Helsinki, Finland
†Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong
Email: abhishek.kumar@helsinki.fi, braudt@ust.hk, sasu.tarkoma@helsinki.fi, pan.hui@helsinki.fi

***Abstract*—The era of pervasive computing has resulted in countless devices that continuously monitor users and their environment, generating an abundance of user behavioural data. Such data may support improving the quality of service, but may also lead to adverse usages such as surveillance and advertisement. In parallel, Artificial Intelligence (AI) systems are being applied to sensitive fields such as healthcare, justice, or human resources, raising multiple concerns on the trustworthiness of such systems. Trust in AI systems is thus intrinsically linked to ethics, including the ethics of algorithms, the ethics of data, or the ethics of practice. In this paper, we formalise the requirements of trustworthy AI systems through an ethics perspective. We specifically focus on the aspects that can be integrated into the design and development of AI systems. After discussing the state of research and the remaining challenges, we show how a concrete use-case in smart cities can benefit from these methods.**

***Index Terms*—Artificial Intelligence, Pervasive Computing, Ethics, Data Fusion, Transparency, Privacy, Fairness, Accountability, Federated Learning**

I. Introduction

Recent technical advances in computing and communication have led to a multiplication of devices embedding computational capabilities, a phenomenon more commonly known as pervasive computing. Such devices constantly monitor and sense users and their environment, producing a vast amount of behavioural data [1]. Emerging technologies such as Augmented Reality heavily rely on continuous video feeds of the users' surroundings [2]. Artificial Intelligence (AI) systems can exploit this data to learn more about users. This data collection and interpretation may be directed towards improving the quality of service, but may also serve other purposes, including surveillance, advertisement, and the algorithmisation of behaviours. AI has started to reach domains with a direct impact on human life, including justice, healthcare, and autonomous driving. In such fields, every decision can have dramatic outcomes, and there is no room for erroneous conclusions. However, AI systems are subject to various distortions which may lead to unfair decisions.

As AI systems vastly depend on data, AI-assisted decisions are only as right as the data provided for training. Such data often reflects existing biases in gender, race, or religion. The resulting AI systems will thus reinforce existing discriminations. Recent examples include recruitment algorithms that, being provided data containing a majority of male applicants, penalise resumes from female applicants [3]. Besides the initial data provided for training, some AI systems rely on a continuous data stream for learning. Such systems may be hijacked by malicious users to alter their original purpose. A famous real-world example is Tay, Microsoft's AI-powered chatbot, which started posting offensive tweets after interacting with Twitter users [4]. In addition to existing bias in data, the usage of AI and data themselves may be questionable, as shown by the scandal of Cambridge Analytica influencing voters based on their Facebook profile [5]. Finally, some of the techniques behind AI such as deep neural networks, boosting, and random forest function as "black boxes", making decisions without exposing the underlying reasons, and preventing their application in the most sensitive domains. It is hard to explain how these models behave, how they produce predictions, and what the influencing variables are. At the other end of the spectrum, white-box models such as linear regression and decision tree solve most of the above questions, at the cost of lower accuracy than black-box models. Such examples raise the question of public trust in often unregulated AI systems.

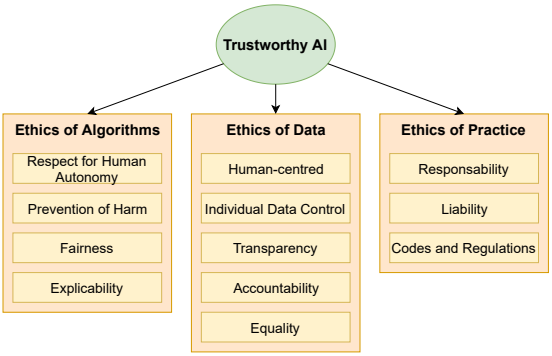

Fig. 1. The three main components of a trustworthy AI [6].

Trustworthy AI can be seen as the sum of three components, shown in Figure 1: ethics of algorithms, ethics of data, and ethics of practice [6]. These components provide a data-centric level of abstractions for ethical questions. Attempting to solve ethical issues for AI systems raises many open issues. What constitutes human ethics can hardly be summarised by typical representations such as decision trees. Transferring human ethics into a clearly defined machine ethics ruleset is thus an intricate problem. In this paper, we target the issues that can be solved at the development stage to provide trustworthy-by-design AI systems. As such, we focus primarily on the ethics of algorithms and ethics of data. The current pervasiveness

of computing resources and recent advances in federated learning allow designing distributed systems addressing many of the current challenges in AI ethics. After discussing the requirements of a trustworthy AI, we extensively review the current state of research on such issues and the remaining challenges. We finally describe how a concrete use-case in smart cities can benefit from such techniques.

II. Ethics of Algorithms

In 2018, the High-Level Expert Group on AI set up by the European Commission released ethical guidelines for building trust in human-centric AI, towards Trustworthy AI [7]. Such an AI system should respect the following ethical principles: 1) Respect for Human Autonomy 2) Prevention of Harm 3) Fairness 4) Explicability. Based on these four principles, the guidelines propose seven core requirements (Figure 2).

A. Requirements for Trustworthy AI

1) Human agency and oversight: AI systems should enforce the *principle of respect for human autonomy*. AI systems should act as enablers to a democratic and equitable society by supporting the user's agency, foster fundamental rights, and allow for human oversight.

2) Technical Robustness and Safety: Technical robustness is related to *prevention of harm*. It requires AI systems to be developed with a preventive approach to risks so that they behave reliably while minimising harm. This should also apply to potential changes in their operating environment or the presence of adversaries attacking the system.

3) Privacy and Data Governance: Privacy is also closely linked to *prevention of harm*, as a fundamental right particularly affected by the pervasive data collection behind AI systems. Prevention of harm to privacy also necessitates data governance that covers the quality and integrity of the data used, its relevance, its access protocols and the capability to process data in a manner that protects privacy.

4) Transparency: This requirement is closely linked to *explicability*. It seeks the transparency of all elements relevant to an AI system: the data, the system and the business models. In the age of pervasive computing, transparency is critical to justify extensive data collection and its benefits to users.

5) Diversity, Non-discrimination and Fairness: In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system. Besides the consideration and involvement of all affected stakeholders, this also entails ensuring equal access and equal treatment. This requirement is closely related to *fairness*.

6) Societal and Environmental Well-being: In line with *fairness and prevention of harm*, the broader society and the environment should also be considered as stakeholders. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concerns. AI systems should benefit all human beings, including future generations.

7) Accountability: The requirement of accountability is closely linked to the *principle of fairness*. It necessitates mechanisms to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.

B. Recent Research Towards Trustworthy AI System

Many research works are tackling the seven requirements mentioned in Section II-A for a trustworthy AI. In this section, we discuss the current status of research and the remaining challenges for each of the seven requirements.

1) Human Agency and Oversight:

Human-in-the-Loop AI: Originally, the Human-in-the-Loop AI approach was proposed as a workflow where AI learns from the human operator while intuitively making the human's work more efficient. Active learning is one such approach where humans provide labels for some unlabelled data to order to achieve the desired accuracy quickly. It can be utilised in designing AI systems. Recently, this approach has been exploited to design fairer paradigms for AI systems, i.e. AI systems which generate revenues will repay the legitimate owners of the knowledge used for taking those decisions [8].

Meaningful Human Control (MHC): In Autonomous Weapon Systems, the MHC paradigm ensures that humans have the power to influence or direct the course of events as well as the ability to manage a machine [9]. Using this paradigm to design AI systems can ensure the capability for human intervention during both the design and the operation, including the possibility to override a decision made by the AI system when it violates the law.

2) Technical Robustness and Safety:

Adversarial Artificial Intelligence: AI models are susceptible to various kinds of attacks, i.e. poisoning attacks [10] [6], evasion attacks [11], and model stealing attacks [12]. Current state-of-the-art defence methods against these attacks focus on increasing the robustness of the model by injecting adversarial examples into the training set [13], hiding the model's information from adversaries [14], reducing the sensitivity of the model by reducing its complexity [15] or minimising the transferability (especially in neural networks) [16].

Safe and Reliable Artificial Intelligence: Reliability in AI systems is ensured by borrowing three principles: 1) Failure Prevention, 2) Failure Identification & Reliability Monitoring, and 3) Maintenance. To prevent failures in AI systems, the current state-of-the-art methods try to proactively identify likely sources of error resulting from 1) bad or inadequate data [17], 2) differences or shifts in environment [18], 3) model associated errors [11], or 4) poor reporting [19] and develop methods that correct for these in advance. After deployment, reliability mechanisms assess the model output for each new input and reject the unreliable output based on the auditing criteria of the density principle and the local fit principle [20]. Model maintenance requires detecting when updates to the model are necessary; however, unlike in traditional software engineering systems, the maintenance cost is already compounded due to development complexity [21].

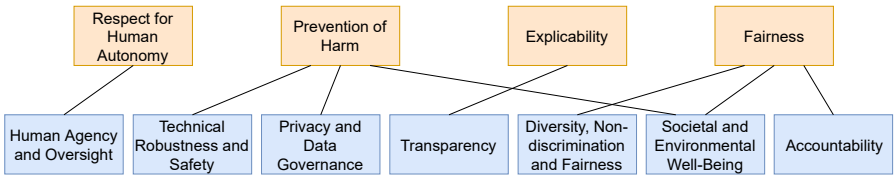

Fig. 2. Ethics of Algorithms [7].

3) Privacy and Data Governance:

Privacy by Design: The privacy by design approach calls for privacy concerns to be predominant throughout the whole engineering process. It was originally designed for traditional software systems. In the context of AI systems, it entails seeking explicit consent from data owners before using their data in training the model and respecting clauses like the "right to be forgotten" under General Data Protection Regulation (GDPR), a regulation in EU law on data protection and privacy in the European Union and the European Economic Area, whenever invoked by the data contributors. Efficient solution for such implementing such clauses in AI systems has just begun to be proposed and largely remains unexplored [22] [23].

Federated Architecture: Federated architectures enable training AI systems without collecting users' personal data at a centralised location. With federated learning [24] or peer-to-peer (P2P) machine learning [25], users do not have to share raw data, rather only training updates. Differential privacy [26] and secure aggregation [27] enforce another layer of privacy to make training updates more secure.

4) Transparency:

Explainable Artificial Intelligence: Explainability in AI systems requires that the decisions made by an AI system can be understood and traced by human beings. Explainability in traditional AI systems (using rule/tree-based models like decision tree, linear models like linear regression, Gaussian mixture model) is straightforward because the decision boundaries corresponding to these models are easy to visualise. Explainability in deep learning AI systems is not possible due to its non-linear structure (hence, known as black-box models). Current attempts of explainability for black-box models are focused on input attribution [28], concept testing/extraction [29], example influence/matching [30], and distillation [31].

Communication: AI-based Chatbots like Google Duplex can mimic human sound so perfectly that even a human could not tell whether they were talking to a robot (may be used for telephone-based scams [32]). Generative adversarial networks can produce fake images which look real to human users [33] [34]. Transparency by communication stresses that AI systems should not present themselves as humans to users and humans have the right to be informed that they are interacting with an AI system. Efficient solutions to distinguish human initiated action from machine initiated action remain unexplored.

5) Diversity, Non-discrimination and Fairness:

Discrimination Discovery: Discovery aims at finding discriminatory patterns in data using data mining and machine learning methods [35] [36]. It builds upon extensive research in statistics on discrimination evidence [37], addressing new challenges due to the increasing volumes and complexity of data and ways of possible unfairness. Statistics has been focusing on hypotheses testing in decision data and provide essential solutions to compare groups of people correctly.

Discrimination Prevention: Discrimination prevention develops methods for sanitising algorithms or adjusting machine learning processes so that outputs obey the fairness constraints. Several attempts to fix algorithms include prepossessing training datasets [38], adding a regularizer to the model [39], post-processing trained models [40] or model outputs [41].

6) Societal and Environmental Well-being:

Computational Sustainability: Computational sustainability requires the development and deployment of AI systems to tackle pressing societal concerns in the most environmentally friendly way possible. It involves efficient resource usage and energy consumption during training and deployment with minimum carbon footprint [42] [43]. In smart cities, AI-based transportation systems plan efficient travel routes while minimising greenhouse gas emissions for commuters [44].

Affective Computing: Ubiquitous exposure to social AI systems due to wide-scale usage of wearables and social media in many of our lives can alter our conception of social agency, or impact our social relationships and attachment. For example, algorithms managing the newsfeed of Facebook users may influence the users' political perception [45]. Affective computing advocates computers with empathy and giving emotional intelligence to machines [46].

7) Accountability:

Data Provenance: Data provenance methods track the flow of data from end to end, across technical and administrative boundaries, thereby bringing accountability by providing evidence—for example, how personal data is collected and subsequently processed, improper behaviour such as unjustified personal data disclosure to an advertiser [47] [48].

Data Auditing: Data auditing for AI systems takes place during both development and deployment. During the development, software verification methods [49] ensure compliance with regulations and company policies. For example, GDPR mandates privacy (and data protection) by design and by default, i.e. seeking explicit consent from users and applying the strictest privacy settings by default. During deployment, blockchain or distributed ledger approaches allow to establish "immutable" records that can operate without requiring trusted third parties and hence maintaining integrity [50].

III. Ethics of Data and Ethics of Practice: Towards Trustworthy AI Systems

AI systems differ from traditional decentralised systems in one significant aspect: AI models revolve around data. It is one of the main reasons why companies with an active AI engagement, such as Google, Amazon, or Facebook, collect users' personal information pervasively. This data collection includes information actively shared by the user, but also a multitude of other parameters sensed by other devices. In return, users benefit from their services for free. This arrangement can be a win-win scenario for both parties. However, some recent abuses threaten the implicit understanding of this arrangement. One may remind the supermarket Target figuring out the pregnancy of a teenager before her father [51], or Cambridge Analytica influencing voters based on their Facebook profile [5]. Such scandals call for a set of measures on data ethics, that is the responsible and sustainable use of data by companies, authorities and organisation to sustain user's trust. DataEthics, a Denmark-based politically independent ThinkDoTank, recommends five principles: **1) Human being at the center, 2) Individual data control, 3) Transparency, 4) Accountability, and 5) Equality** [52] to enforce data ethics.

A. Ethics of Data

The ethics of data focuses on ethical problems posed by the collection and analysis of large datasets using AI or Data Mining techniques which makes it possible to re-identify individuals via linking with other auxiliary datasets. With the advent of 5G, sensing and AI are becoming more pervasive, supported by edge computing, in-network AI, augmented reality, and the Internet of Things [53]. This phenomenon leads to the problem of *ethics of sensing*. Where can we place the logic that guides and governs the ethical behaviour of a system or application? Should some of the logic be in the sensor itself to prevent misuse? Two directions can enforce data ethics. A first direction lies in using design principles which advocate storing very little or virtually no amount of user's raw data. The other direction is letting users decide which data they want to share and even earn economic benefits [54]. This approach aims to give control of personal data back to users. It can play a significant role in challenging the current status-quo of *Surveillance Capitalism*, i.e. the commodification of users' personal data and their transformation into behavioural data for analysis and sale.

1) Design Principles for Ethics of Data:

- **Zero Knowledge as a Design Principle:** According to the GDPR, no data must be stored longer than is necessary. Companies can go beyond the legislation and delete data before the required date. It can be done through auto-deletion, but also by never having access to the data in the first place. Federated learning or P2P based AI systems are an example of such design principles where the user's raw data never leaves their device [24].
- **Contextual Integrity as a Design Principle:** Many organisations collect data on people's lives and activities without their knowledge [55]. In 2016, a group of Danish researchers publicly released a dataset of nearly 70,000 users of the online dating site OkCupid, including usernames, age, gender, location, sexual orientation, personality traits, and answers to thousands of personal profiling questions [56]. Although such information was already public on OkCupid, users did not intend for a study to exploit their data. The Facebook–Cambridge Analytica data scandal is another emblematic example. The theory of Contextual Integrity allows enforcing data ethics by providing a framework for evaluating the flow of personal information between agents to identify and explain why certain patterns of information flow are acceptable in one context but viewed as problematic in another [57] [58].

2) Personal Data Infrastructure for Ethics of Data:

- **Personal Databox:** The idea of a personal databox was proposed by Chaudhry et. al in 2015 [59]. Personal databoxes act as a single-point (physical) gateway through which users' data flows to anywhere outside the user's control. It allows users to capture, index, store and manage data about them as well as data generated by themselves.
- **Hub-of-All-Things (HAT):** HAT provides users with a personal data infrastructure. Unlike Databox, HAT relies on the idea that users should earn economic value from their data [60]. Users generate durable data and perishable data. Perishable data is often not used within its expiry window and hence, loses its value. HAT allows capturing such data and brokers benefits from their parties available in its ecosystem. HAT advocates the fact that users should receive explicit benefits for their data (privacy trading).

B. Ethics of Practice

Ethics of practices focuses on the responsibilities and liabilities of people and organisations in charge of data. It aims to determine and maintain guarantees for the ethical use of data and models when disclosed at multiple stages and aggregated by multiple parties. To this purpose, it is necessary to define both a technical [61] and ethical framework to guide decisions such as data release without getting into scandals like the AOL search data leak [62] or Netflix privacy lawsuit [63]. More recently, it has also extended to the practice of model and synthetic model disclosure, since models trained on user's data may release private information using new types of attacks like membership inference attack [64].

IV. Use-case: Trustworthy AI in Smart Cities

Smart cities are a concrete example of pervasive computing and data sensing. A multitude of sensors collect data for AI models to provide insights on traffic management and road safety [65], infrastructure monitoring, or community service planning. However, such data collection will inherently expose users' private data to a risk of misuse by authorities and companies, such as location profiling.

Consider the following scenario: *"A city council plans to install chargers for electric vehicles and needs spatiotemporal*

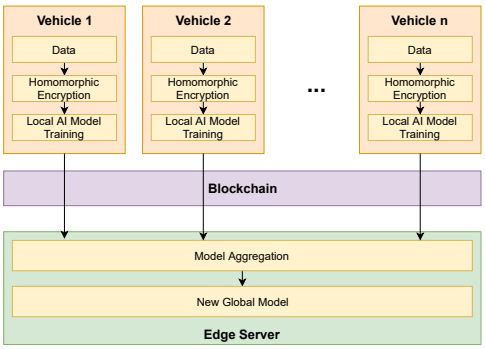

Fig. 3. Trustworthy AI in a smart city scenario.

information from citizens' vehicles. The number and the location of the chargers will depend on the collected information. Citizens are invited to participate in the data collection process." The city council has two options: 1) collect and anonymise citizens' data 2) build an AI system which will recommend the number and the locations of the installations.

Recent advances in data mining techniques have made deanonymisation obsolete. For example, De Montjoye et al. showed that four spatiotemporal points are enough to uniquely identify 95% of people in a mobile phone database of 1.5M people [66]. Building an AI system based on on-device learning, such as federated learning, protects against these problems since the data never leaves the vehicle. Given that vehicles have enough computing power, using federated learning in smart cities' vehicular networks for predicting the number and the locations of the charger installations is technically feasible while protecting users' privacy. Since federated learning systems rely on a zero-knowledge design, they can address the *ethics of data* by on-device processing, i.e. raw data of users do not leave their device. An additional layer of privacy can be enforced with differential privacy or homomorphic encryption in the user's training algorithm. Such a system combined with a distributed ledger such as blockchain allows for accountability as each model update gets logged as an immutable record (as shown in Figure 3). With the increasing number of users and contributors, the consensus mechanism responsible for updating the global model will become computationally expensive, thus costing more energy and preventing real-time applications. A potential solution lies within adding an extra layer to the Blockchain to address the scalability issue. For instance, [67] adds a Management hub to the Blockchain network to tackle access control problem in the presence of billions of IoT devices. Finally, the distributed nature of federated learning allows for resilience to system failure. On the other hand, the architecture alone cannot address all the requirements. Fairness, explainability, and resilience against evasion attacks can be achieved through AI algorithms that tackle the black-box nature of current state-of-the-art AI algorithms, bringing more transparency to the system.

V. Conclusion

In this paper, we proposed a framework for trustworthy AI systems in the age of pervasive data collection and computing. This framework provides a data-centric level of abstractions for ethical questions posed in the AI and Data Science context. This Data-centric level of abstractions provides ethical abstractions on there level: data, algorithm, and practice. We focused on the ethics of data and ethics of algorithms, and more specifically, the aspects that can be integrated directly within the design and development of AI systems. After reviewing the challenges and requirement, as well as the current status of research, we discussed how a concrete use-case in the context of smart cities could benefit from the existing techniques for a more trustworthy usage of AI.

Acknowledgment

This research has been supported in part by project 16214817 from the Hong Kong Research Grants Council, the 5GEAR project and the FIT project from the Academy of Finland.

References

[1] H. Haddadi, P. Hui, T. Henderson, and I. Brown, "Targeted advertising on the handset: Privacy and security challenges," in *Pervasive Advertising*. Springer, 2011, pp. 119–137.

[2] C. Bermejo, Z. Huang, T. Braud, and P. Hui, "When augmented reality meets big data," in *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, June 2017, pp. 169–174.

[3] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," *Reuters*, 2018.

[4] S. Perez, "Microsoft silences its new a.i. bot tay, after twitter users teach it racism," *techcrunch.com*, 2016.

[5] "Cambridge analytica and facebook: The scandal and the fallout so far," *The NewYork Times*, 2018.

[6] L. Floridi and M. Taddeo, "What is data ethics?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2083, 2016.

[7] A. Hleg, "Ethics guidelines for trustworthy ai," *B-1049 Brussels*, 2019.

[8] F. M. Zanzotto, "Human-in-the-loop artificial intelligence," *Journal of Artificial Intelligence Research*, vol. 64, pp. 243–252, 2019.

[9] I. Verdiesen, "The design of human oversight in autonomous weapon systems," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2018, pp. 388–389.

[10] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," in *Asian Conference on Machine Learning*, 2011, pp. 97–112.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[12] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," in *25th {USENIX} Security Symposium ({USENIX} Security 16)*, 2016, pp. 601–618.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[14] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3*, 2018.

[15] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*, 2018.

[16] H. Hosseini, Y. Chen, S. Kannan, B. Zhang, and R. Poovendran, "Blocking transferability of adversarial examples in black-box learning systems," *arXiv preprint arXiv:1703.04318*, 2017.

[17] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, 2018, pp. 77–91.

[18] A. Subbaswamy, P. Schulam, and S. Saria, "Preventing failures due to dataset shift: Learning predictive models that transport," in *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, 2019.

[19] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31,*, 2019.

[20] P. Schulam and S. Saria, "Can you trust this prediction? auditing pointwise reliability after learning," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 1022–1031.

[21] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young, "Machine learning: The high interest credit card of technical debt," in *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, 2014.

[22] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symposium on Security and Privacy*, 2015.

[23] D. Barua, "A time to remember, a time to forget: user controlled, scalable, life long user modelling," *The University of Sydney, Australia*, 2016.

[24] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[25] A. Bellet, R. Guerraoui, M. Taziki, and M. Tommasi, "Personalized and private peer-to-peer machine learning," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 473–481.

[26] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[27] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2017.

[28] A. Kapishnikov, T. Bolukbasi, F. B. Viégas, and M. Terry, "Segment integrated gradients: Better attributions through regions," *CoRR*, vol. abs/1906.02825, 2019.

[29] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Advances in Neural Information Processing Systems*, 2019.

[30] S. Ö. Arik and T. Pfister, "Attention-based prototypical learning towards interpretable, confident and robust deep neural networks," *CoRR*, vol. abs/1902.06292, 2019.

[31] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[32] H. Li, X. Xu, C. Liu, T. Ren, K. Wu, X. Cao, W. Zhang, Y. Yu, and D. Song, "A machine learning approach to prevent malicious calls over telephony networks," in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 53–69.

[33] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[34] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[35] S. Ruggieri, D. Pedreschi, and F. Turini, "Data mining for discrimination discovery," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 4, no. 2, p. 9, 2010.

[36] K. Mancuhan and C. Clifton, "Combating discrimination using bayesian networks," *Artificial intelligence and law*, vol. 22, no. 2, 2014.

[37] T. Tinkham, "The uses and misuses of statistical proof in age discrimination claims," *Hofstra Labor and Emploment Law Journal*, vol. 27, 2010.

[38] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 259–268.

[39] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[40] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, July 2013.

[41] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 7, 2012.

[42] D. H. Fisher, "A selected summary of ai for computational sustainability," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[43] J. Khakurel, B. Penzenstadler, J. Porras, A. Knutas, and W. Zhang, "The rise of artificial intelligence under the lens of sustainability," *Technologies*, vol. 6, no. 4, p. 100, 2018.

[44] J. A. Guerrero-Ibanez, S. Zeadally, and J. Contreras-Castillo, "Integration challenges of intelligent transportation systems with connected vehicle, cloud computing, and internet of things technologies," *IEEE Wireless Communications*, vol. 22, no. 6, pp. 122–128, 2015.

[45] O. Alvarado and A. Waern, "Towards algorithmic experience: Initial efforts for social media contexts," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018.

[46] R. W. Picard, *Affective computing*, 2000.

[47] J. Singh, J. Cobbe, and C. Norval, "Decision provenance: Harnessing data flow for accountable systems," *IEEE Access*, vol. 7, 2018.

[48] T. Pasquier, J. Singh, J. Powles, D. Eyers, M. Seltzer, and J. Bacon, "Data provenance to audit compliance with privacy policy in the internet of things," *Personal and Ubiquitous Computing*, vol. 22, no. 2, pp. 333–344, Apr 2018.

[49] D. R. Wallace and R. U. Fujii, "Software verification and validation: an overview," *Ieee Software*, vol. 6, no. 3, pp. 10–17, 1989.

[50] J. Dai and M. A. Vasarhelyi, "Toward blockchain-based accounting and assurance," *Journal of Information Systems*, vol. 31, no. 3, 2017.

[51] K. Hill, "How target figured out a teen girl was pregnant before her father did," *forbes.com*, 2012.

[52] DataEthics, "Dataethics – principles and guidelines for companies, authorities & organisations," pp. 1–38.

[53] F. B. Saghezchi, J. Rodriguez, S. Mumtaz, A. Radwan, W. C. Lee, B. Ai, M. T. Islam, S. Akl, and A.-E. M. Taha, "Drivers for 5g: The 'pervasive connected world," *Fundamentals of 5G Mobile Networks*, pp. 1–27, 2015.

[54] K. C. Laudon, "Markets and privacy," *Communications of the ACM*, vol. 39, no. 9, pp. 92–104, 1996.

[55] M. Zimmer and K. Kinder-Kurlanda, *Internet research ethics for the social age: New challenges, cases, and contexts*. Peter Lang International Academic Publishers, 2017.

[56] R. Hackett, "Researchers caused an uproar by publishing data from 70,000 okcupid users," *Fortune*, 2016.

[57] H. Nissenbaum, "Privacy as contextual integrity," *Wash. L. Rev.*, vol. 79, p. 119, 2004.

[58] M. Zimmer, "Addressing conceptual gaps in big data research ethics: An application of contextual integrity," *Social Media + Society*, vol. 4, no. 2, 2018.

[59] A. Chaudhry, J. Crowcroft, H. Howard, A. Madhavapeddy, R. Mortier, H. Haddadi, and D. McAuley, "Personal data: thinking inside the box," in *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives, 2015, Aarhus, Denmark, August 17-21*, 2015.

[60] I. C. Ng, "The market for person-controlled personal data with the hub-of-all-things (hat)," 2018.

[61] G. Zhang, T. Li, Y. Li, P. Hui, and D. Jin, "Blockchain-based data sharing system for ai-powered network operations," *Journal of Communications and Information Networks*, vol. 3, no. 3, pp. 1–8, 2018.

[62] C. Chiru, "Search engines: Ethical implications." *Economics, Management & Financial Markets*, vol. 11, no. 1, 2016.

[63] R. Singel, "Netflix cancels recommendation contest after privacy lawsuit," *WIRED Magazine*, 2010.

[64] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, May 2017, pp. 3–18.

[65] P. Zhou, T. Braud, A. Alhilal, P. Hui, and J. Kangasharju, "Erl: Edge based reinforcement learning for optimized urban traffic light control," in *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, March 2019.

[66] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 2013.

[67] O. Novo, "Blockchain meets iot: An architecture for scalable access management in iot," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1184–1195, April 2018.