Consumer Artificial Intelligence Mishaps and Mitigation Strategies

Sirwe Saeedi
Western Michigan University

Saraju P. Mohanty
University of North Texas

Steve Carr
Western Michigan University

Alvis C. M. Fong
Western Michigan University

Ajay K. Gupta
Western Michigan University

***Abstract*—Although artificial intelligence (AI) promises to deliver ever more user-friendly consumer applications, recent mishaps involving fake information and biased treatment serve as vivid reminders of the pitfalls of AI. AI can harbor latent biases and flaws that can cause harm in diverse and unexpected ways. Before AI becomes interwoven into human society, it is important to understand how and when AI can fail. This article presents a timely survey of AI-induced mishaps that relate to consumer applications. The article also offers suggestions on mitigating strategies to manage the undesirable side effects of using AI for consumer applications. It therefore serves a dual purpose of creating awareness of current issues and encouraging other researchers in the consumer technology (CT) community to build better AI consumer applications.**

Artificial intelligence (AI) powers a wide range of smart consumer devices and applications [1]. Machine learning (ML), which grew out of AI, has been a major driver of recent AI advances. From consumer imaging systems [2] to home safety [3] and personal stress monitoring [4]-[5], consumer AI technology permeates everyday life. While AI can empower user-friendly applications, the outcome can be unpredictable, e.g., face misidentification due to biases. There are vulnerabilities associated with the "black box" nature of some ML algorithms underpinning AI that can harbor latent biases that are potentially harmful to consumers.

Despite years of development in advanced deep neural networks (DNNs), researchers are still improving their understanding of how DNNs operate. End users and other stakeholders (e.g. data curator) have a part to play because their technical understanding is often limited and dangerously prone to anthropomorphic tendencies that can be replicated or even amplified algorithmically.

The vulnerabilities of ML algorithms [6]-[8] include (see Figure 1): (1) data dependency, i.e.,

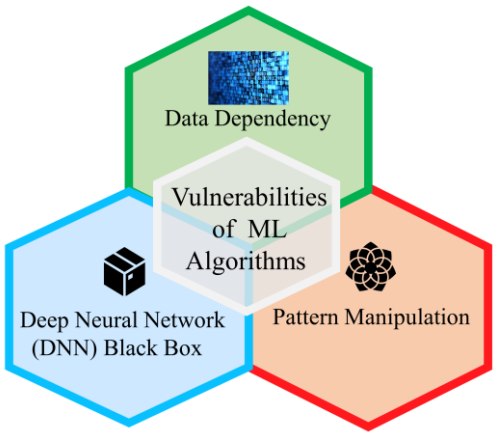

Figure 1. Vulnerability of AI/ML.

algorithms' reliance on data, which may be biased, incomplete, or defective, (2) learning statistical patterns that are easy to manipulate, and (3) the black box nature of contemporary DNNs means it is not always clear how decisions are made, which can perpetuate or hide biases. Together, these characteristics explain why vulnerabilities can be targeted by adversaries or triggered unintentionally.

In fact, the main problem lies in dataset bias, and ML models tend to perpetuate inherent flaws in the data. In ML, the (training) dataset is all that an algorithm sees; the dataset is the world [9]. A biased dataset is one that instead of training a model to have the ability to generalize in the real open world, the ML model becomes a closed world [9]. An example of dataset bias is the following: if a cow frequently appears together with grass in the training data, then detecting grass and outputting "cow" can become a characteristic of the resultant ML classifier [10]. Further, such biases tend not only be replicated by ML but worsened through bias amplification [11].

Discrimination can result from biased data, causing some people to feel unfairly treated. When an ML algorithm focuses on the majority group in a dataset while accepting high error rates for minority groups, it can lead to amplification of existing disparities [10], [12]. This can even generate new disparities over time [13].

Contemporary DNNs tend to obscure how decisions are made, so flaws become even harder to detect. Incomplete or defective data can often lead to biases. ML algorithms learn from data (examples) presented to them through a training process. Once training is complete, the ML model is deployed to work on new, unseen data. For example, if a face recognition algorithm is trained on data that predominantly comprise images of faces of certain demography, the training data are incomplete in the sense that these faces do not represent the general population. ML, otherwise known as statistical learning, will learn from the incomplete/biased data to recognize these faces when the model is deployed. These ML models are statistically impressive (achieving good overall performance), but individually unreliable. The statistical nature of these models amplifies algorithmic bias. All these biases often manifest themselves in discrimination when these models are deployed in the real world. For example, the face recognizer may show poor performance when tested on facial images of other demographic groups. Many best performing facial recognizers are built on DNNs, making detection of biases difficult. A study found that four popular face recognizers performed better one gender than another and better on one color faces than another [14]. The worst of the four had a 34% error rate [14].

It is critical to distinguish between cyberattacks and AI vulnerabilities [15], [16]. Cyberattacks are deliberate exploitation to gain unauthorized access then target infrastructures. Unlike cyberattacks, AI mishaps are often caused by inherent vulnerabilities of ML. Cyberattacks need sophisticated techniques, but bad actors with limited technical knowledge can use AI to deceive others. For example, Deepfake allows unskilled people to fabricate fake texts, images, or videos using consumer grade AI tools, regardless of the complexity underlying their algorithms [17], [18]. These fake artifacts can be harmful and mislead consumers [53].

This article will present AI-induced mishaps that relate to consumer applications. Although the list of ethical challenges in the complex field of AI may be prohibitive, this article aims to create awareness of the issues. The topic of adversarial attacks is best covered under cybersecurity and is not further discussed in this article. Instead, this article focuses on hidden biases and touches on deepfakes to highlight considerable risks of AI technologies on segments of consumers.

BIASES AND DISCRIMINATION

The cognitive bias is prevalent (see Figure 2). People may have gender bias for certain types of jobs [19], [20]. These biases are often coded into data and learned by algorithms [21]-[24]. Debiasing can be expensive, time consuming, and may be impossible. So, profit-driven companies often overlook such biases, which can be unfair to consumers. Mishaps in consumer applications reveal more hidden discrimination patterns in data against human diversity in the deepest layers of DNNs.

Online Platforms

Online platforms support many daily activities but can be a source of biases (see Figure 2). A commonly applied ML technique in the online platforms is predictive models. Since data are often biased, ML predictive analytics may reflect undesirable decisions and perpetuate biases. Online advertising technology can negatively impact certain demography [25]. Specifically, the delivery of the largest provider of online delivery advertisements is statistically discriminatory based on names typically associated with certain communities. For example, when someone searches for a person named "XXX", an advertisement that suggests XXX has a criminal record might pop up alongside XXX's list of accomplishments. A side effect of this bias is worsening algorithm's performance by frequently selecting those advertisements.

Ride hailing service platforms have also been found to exhibit algorithmic bias. A study analyzed 100 million rides found that major ride-hailing companies had unfair charges for certain neighborhoods [28]. Online purchase and delivery systems also show biases [29]. Although these services claim that they do not differentiate consumers' ZIP codes, significant differences have been found in the availability of one-day delivery of consumer products in different neighborhoods.

Employment Recommendation Systems

Recommendation engines could improve processing of job applicants and even predict future preferences. An AI-empowered model may not fairly rate applicants for vacancies [30]. It underestimated résumés of applicants one gender because of limitations in ML techniques that are mainly trained by résumés from another gender.

Another example found that certain names enhances the chance of success in hiring for no clear reasons [31]. Another failure of hiring algorithm shows a strong correlation between a variable of the model (commuting distance) and certain demography [32]. Apart from defective data, statistical predictions tend to make decisions like what is recommended [33]. Another concern with predictive hiring systems is fairness for people with disabilities. An example of excluding applicants using assistive technologies like magnifier or screen reader, even though their disabilities were not mentioned [34].

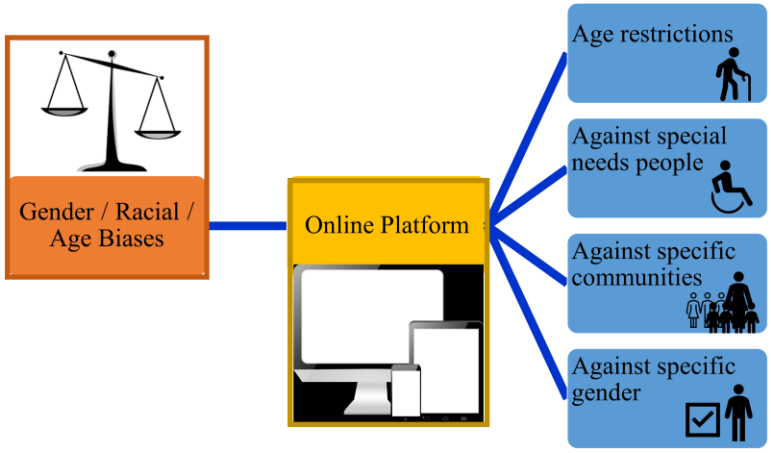

Figure 2. AI/ML biases and discrimination.

Natural Language Processing Systems

Natural Language Processing (NLP) is crucial to a wide range of voice activated consumer products [35]. NLP is applied to understand patterns in the unstructured data like text and voice with all hidden or plainly visible stereotypes. Word embedding tools, an underlying task in NLP, represent a word by vectors of trying to represent the true meaning. ML models can learn associations between concepts like female names with family and male names with professional jobs [36].

AI-enabled NLP biases extend well beyond gender biases. A research group analyzed millions of tweets on a popular messaging platform containing using NLP tools. They provided quantitative problematic evidence of demographic bias in classifying tweets [37]. In the NLP domain, a slightly perceptible manipulation can change the semantics and syntax of text. Robustness of DNNs on sentiment analysis and textual entailment tasks can be disturbed easily by the generated adversarial examples [38].

This problem extends to any AI system that uses NLP, including voice. A study shows how a speech-to-text engine is deceived by adding a small distortion to voice and turning the original voice to

target transcription [39]. Some other AI systems are biased against strong or uncommon dialects. Voice recognition systems are supported by NLP models and they may not learn diverse data. A study has found that two popular commercial voice interfaces misidentify voices of certain demography 35% of the time [40].

A known mishap involving AI chatbot occurred in 2016 when it was taken down after only a few hours due to offensive tweets [41]. The bot learned from the tweets and searched the internet to find data for its responses.

Thus, an important question is: can we trust machines that generate human-like output and make human-like decisions? As AI/ML models get better in understanding patterns of human culture, new challenges of weaponizing NLP tools emerge, such as generating misleading propaganda, fake content production, phishing emails attacks, and impersonating other users.

Predictive Tools

AI predictive tools have been flagged as a threat to customer privacy and fair treatment (see Figure 3). For example, a company made sensitive decision about female customers' pregnancy status [42]. When a father saw the company's coupons for baby items, he realized his daughter was pregnant. The main problem is where predicted sensitive information can cause erosion of privacy and trust if not used appropriately.

In another study, a family screening tool designed to improve child welfare was found to be acting on prejudiced data [43]. Analysis of phone calls to hotline unreasonably reported families of a specific demography to be suspected more often than others. In addition to privacy concerns, predictive tools have been reported to cause harm to consumers seeking loans and healthcare services.

Fairness in Consumer Lending

AI-driven lending tools can provide both advantages and disadvantages. Making better consumer lending decision needs to update variables to extract patterns that indicate creditworthiness. A recent report that evaluated the impact of AI technology in consumer lending and claims that US regulatory structure could not guarantee to protect fair lending foundations against different types of discrimination [44].

Credit reporting bureaus use metrics like income and credit scores that are correlated with gender, race, and other demographic attributes. An analysis shows the average credit scores of homeowners of a specific demography are substantially higher than another [44]. Another study concludes that US credit scoring systems amplify demographic disparities because it is the most important criterion considered by financial companies [45].

Many studies show credit-based insurance mechanisms are biased against specific demography [46],[47]. The single predictive variable that had a direct impact on reporting insurance score and premium was demography. Automatic background check systems provide homeowners with a single score to determine the eligibility of tenants [48]. The screening tools to predict the potential risk show bias against specific neighborhoods.

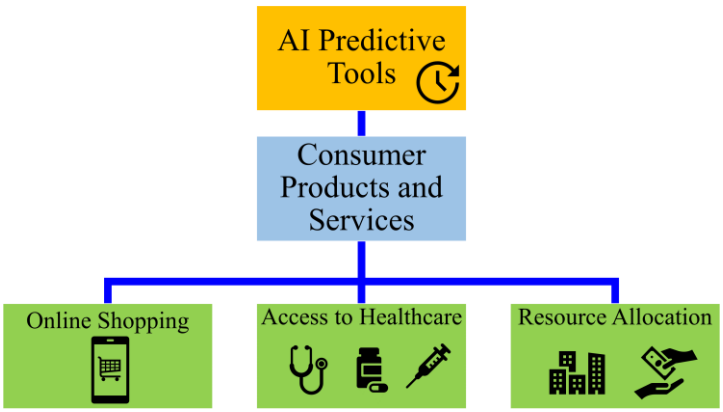

Figure 3. AI predictive tools.

Fairness in Consumer Healthcare

The healthcare sector is turning to AI to help people in need of medical care. However, mismanaged tools in a mission-critical area can have huge implications on human lives. An unjust AI system could target residents in a low-income neighborhood with serious illness and disorders. There are many instances that support the imperfections and injustices in intelligent

healthcare. In one case, a woman with cerebral palsy lost her healthcare plan without any explanation from the providers; the algorithm falsely recognized her as a non-emergency patient [49].

Developers of an automated healthcare system implemented more than 900 unfair rules into the model code, resulting in mistakenly deleting patients in desperate need [50]. A predicting tool used to assess patient situation with pneumonia made a serious mistake [51]. The algorithm persuaded doctors to send patients home despite their extensive medical problems.

A dataset lacking patients with diverse medical backgrounds could worsen health disparities in the model. For example, a multivariable linear regression model employed to assess cardiovascular risk score using data derived from almost exclusively people of a specific demography is less accurate among different other groups [52]. The automated medical tools may recommend no further treatment for cases ignored by machine.

CONTENT CREATION AND FILTERING

Ready availability of AI tools has lowered the barrier for non-experts to create fake content. With the proliferation of fake content, effective countermeasures are needed for consumers to protect themselves from harm. However, content filters intended to reduce the effects of information overload can also adversely affect consumers.

Content Creation – Deepfake

Deepfake uses AI to generate realistic video or audio content designed to deceive [53]. Instances of fake content abound. For example, a group of researchers transformed audio clips of a former US president into a lip-synced video clip [54]. The system has the potential to put other people's words into someone's mouth. Further, Deepfake can create non-existent unique faces to mimic a real person. A bot that generated Deepfake text from real submissions for a federal public comment website was so convincing that even a human classifier was not better than random guessing in discerning bot submissions from real comments [17].

Content Filters

Social networks have surpassed newspapers as primary news outlets for many; trending topics represent popular news. Therefore, one of the critical areas affected by AI systems is broad spectrum of content through social networks. Content control software is a part of digital immune systems. Content-based recommendation systems use AI content filtering algorithms to suggest topics related to a user's interested area, based on previous feedback. The type of content that can pass through can have serious consequences.

For example, social network companies have developed their own censorship for enhancing benefits for their users and security. A leading video sharing platform can show more than 700 million hours of video every day, and a popular messaging platform can process up to 500 million tweets each day. Because of the high volumes, they are susceptible to misuse, such as promotion of misinformation, polarization, and violence.

There is often a delicate tradeoff between filtering too much or too little potentially harmful content. In 2017, a popular video sharing platform aggressively deleted more than 31 million videos predicted to include violent content. However, it was found that educational and legitimate documentary videos were deleted mistakenly [55]. The search for an optimal balance between too much and too little filtering remains an open research question. The situation is further complicated by social network companies' vested interests. Algorithmic decisions on what content to recommend or remove are often aligned with maximum engagement not facts [56].

POTENTIAL SOLUTIONS

AI-induced mishaps trace their root causes to three fundamental limitations of contemporary ML techniques. Statistical ML has been the main driver of recent advances in AI. ML has enabled a wide range of consumer applications [57].

Key Limitations of Contemporary Techniques

ML has been the main driver of recent AI advances. But a major limitation of ML algorithms is that they learn from training data before they are validated and deployed. Learning algorithms are designed to effectively learn the nuances in the training data. Thus, any inherent biases will be

learned, and often these biases become entrenched, reinforced, and amplified as learning continues.

A second limitation of statistical machine learning, which characterizes contemporary cutting-edge AI, is attributed to statistics. These ML algorithms, including those developed for DNNs, are statistically brilliant but individually unreliable (see Figure 4). For example, what does it mean when a method can achieve 99% accuracy in binary classification?

In the context of some computer vision tasks, such as recognizing images for non-critical use, 99% accuracy is arguably better than most humans. For some safety critical applications, even this level of performance can be unimpressive. For example, if an AI-powered robotic surgeon with 99% accuracy has performed 99 successful operations, it does not necessarily imply the next patient is doomed. Statistically, 1 in every 100 patients will be adversely affected. However, it is important to compare statistical performances against what is achievable without AI.

The failure mode of the best techniques is not well understood, which leads to the third major limitation. The very best performing DNNs of today are often treated as black boxes. These models are typically hundreds of layers deep and have millions or even billions of parameters; they are often too complex for researchers and practitioners to fully understand their behaviors.

Mitigation Techniques

The first step in mitigation is to create awareness among developers and users about the limitations of contemporary AI and not necessarily rely on it blindly. For example, awareness of biases can help to mitigate the problems [26],[31].

Given that the most prevalent root cause of reported mishaps is attributed to incomplete or biased data (see Figure 5), researchers and developers of AI-enabled consumer devices and applications should exercise caution in data curation. Indeed, Predictive tools can produce and even amplify preexisting bias, technical bias, and emergent bias [58]. Preexisting bias can be traced to entrenched social norms, beliefs, practices, and attitudes. Technical bias results from technical constraints of considerations. Emergent bias occurs in a context of application, such as when a trained

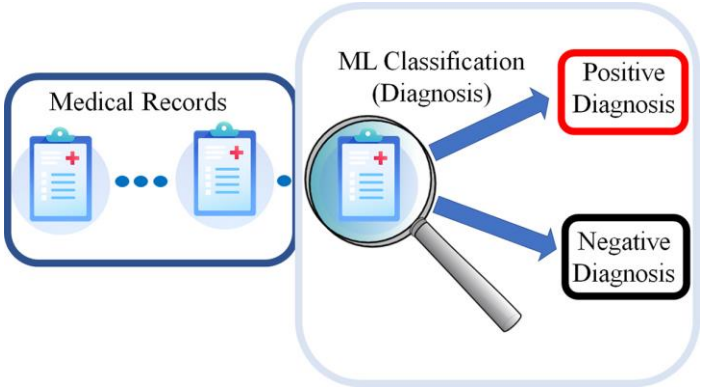

Figure 4. ML binary classification scenario.

ML model is deployed. To mitigate, developers need to be sensitive to possible biases in the data and take corrective action. As a minimum, they need to curate data that represent a broad spectrum of democratic attributes of the intended users. In some cases, eliminating biases from data may be achievable (e.g., through proper sampling and balancing to handle data imbalance, or by eliminating sensitive variables). However, biases are sometimes deeply entrenched in the data, making debiasing a difficult task that remains a subject of intensive research.

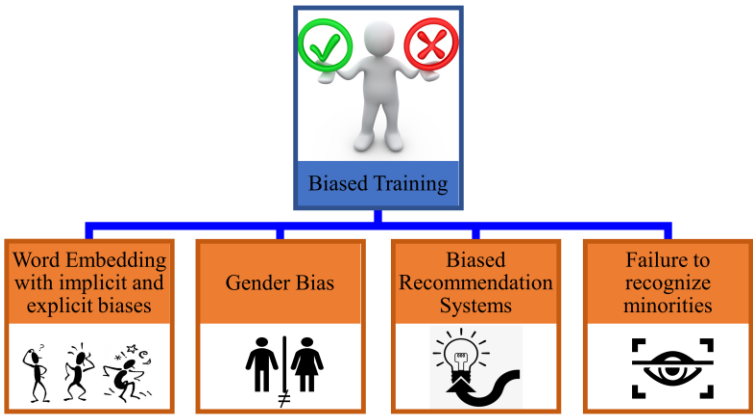

Figure 5. Multi-scenario ill effects of biased training in AI/ML.

Confounding of features can amplify preexisting latent biases. Since AI hiring algorithms absorb many social patterns that reflect demographic discrimination, blindly using them can exacerbate

institutional and systemic biases [30].

Biases in training data are only one aspect of data deficiency. Sometimes, the training data are simply incomplete. As a result, the ML algorithm can persuade doctors to send patients home regardless of their extensive medical problems [51]. If patient conditions are not included in dataset which the system learns from, algorithms could lead to faulty decisions by denying healthcare to needy patients. Designers must therefore consider all kinds of scenarios, even if they are rare, to ensure data completeness.

There are initiatives aimed at enhancing data quality to promote algorithmic fairness, e.g., Data for Democracy [59]. The AI Index is an effort to track, collate, distill, and visualize data for AI [60]. It aspires to be a resource for policymakers, researchers, executives, and journalists, to develop intuitions about the field of AI. The Datasheets for Datasets initiative aims to facilitate communication between dataset creators and dataset consumers for transparency and accountability [61].

In addition to defective data, whether biased or incomplete, predictions with statistical analysis can influence a model's output. For example, statistics dictate that an AI recruiter's decisions will tend to be like what the tool has recommended [33]. Although the most effective mitigating technique or strategy often depends on the root cause, multiple techniques can be applied synergistically. Conversely, one strategy might address two or more issues together. The current effective mitigating strategies for limitations associated with defective data and nature of statistical analysis center around diversification of data for statistical learning. For example, it is argued that an effective way to overcome inequality in medicine is to significantly diversify data [62]. Some experts suggest that debiasing human is harder than debiasing AI and yet the data collection part often requires human input, such as annotation of images for training [63].

The main reasons for the difficulty in weeding out AI biases include [64]: (1) unknown unknowns (effects of biases are felt downstream from where they started), (2) imperfect models (NNs are not typically designed with bias mitigation in mind), (3) lack of social context (social impacts are often not well understood by AI system designers), and (4) notion of fairness is not well understood.

A promising mitigating strategy therefore calls for a lessened reliance on statistical techniques and an elevated involvement of other dimensions of AI. For example, reasoning and abstraction capabilities of AI, such as commonsense reasoning can provide added assurances of the final algorithmic output that was optimized statistically [65], [66].

There is a recent initiative aimed at synergizing good old symbolic AI and connectionist networks. The Neuro-Symbolic (NS) AI initiative aims to address a gap in contemporary AI by leveraging the capabilities of current state-of-the-art statistical ML and classical symbolic AI [67]. A main goal of the initiative is to advance AI to the next level, towards artificial general intelligence (AGI). A useful outcome of this research will be improved understanding of the existing black box methods. The mitigation of bias requires active involvement of AI practitioners and policy makers [68].

While the outcome of this synergistic direction of research is expected in the future, a more immediate mitigating strategy for addressing the black box concern is to strengthen researchers' understanding of how hyperparameter tuning can affect the outcome of opaque DNN. Removal or addition of variables can affect a fairness metric but will not remove embedded bias depending on the robustness of ML models to hyperparameter settings [44]. Therefore, improving transparency in model tuning and hyperparameter settings can lead to enhanced performances.

Another interesting initiative is a "fairness gym", which models fairness as dynamic, and is aimed at understanding long-term fairness [69]. Other technical developments include: (1) causal modeling and counterfactual fairness, (2) bias discovery through fairness aware data mining, and (3) learning latent structures. For example, the ability to reason about counterfactual, what-if scenarios is crucial in the quest to disentangle social biases from the actual phenomenon being modeled [70], [71]. A framework for modeling fairness using tools from causal inference has been proposed [72]. The definition of counterfactual fairness captures the intuition that a decision is fair towards an individual if it is the same in: (1) the actual world and (2) a counterfactual world where the individual

belonged to a different demographic group. Algorithms for discrimination discovery and discrimination prevention with fairness-aware data mining is available [73]. To mitigate gender and racial bias in facial recognition, the original ML task has been combined with a variational autoencoder (VAE) to learn the latent structure in data and then use the learned latent distributions to re-weight the importance of data points while training [74]. An autoencoder takes a high dimensional feature space and compresses it into an encoded (or latent) space characterized by having a lower number of dimensions than the original. A VAE is regularized to minimize overfitting to ensure the latent space will preserve important information on the data points to facilitate reweighting.

The proliferation of consumer grade AI tools for creating fake content and misinformation has made it easy for bad actors to participate in harmful activities. Technical solutions tend to revolve around using similar tools to content creation for detecting fake content. It appears that efforts aimed at creation and detection of fake content are locked in a long-term technical contest. While protection of consumers from harmful fake content remains an active area of research, currently the most effective mitigating tools seem to be based on legislation more than on technical solutions. For example, prohibiting the distribution of fake videos and image targeting high-valued contents.

Conclusion and Future Directions

This article presented a wide range of consumer-impacted AI-related mishaps: personnel recruitment, NLP and voice recognition for interacting with smart devices and assistants, chatbot on smartphones, online and mobile shopping, pop up ads, face recognition, consumer lending, consumer healthcare, social media, and fake and harmful media content. The mishaps revolve around biases, discrimination, and other unfairness due to demographic attributes. Developers of consumer AI applications and products should consider the consequential harmful effects and take steps to avoid them.

Future research is needed to measure the impact of AI mishaps on the CE industry and consumers. Further investigation should be pursued to measure such impact along multiple dimensions, such as financial implications, product design cycle, and consumer protection. The proposed work ties in with the idea that minimizing AI mishaps should be an integral part of the design process and should be considered with end users in mind. Security-by-Design (SbD) principle that advocates to consider cybersecurity as an objective right at the early stage of design cycle can also play a role in designing robust smart electronics design [75].

Acknowledgments

This work was supported by the U.S. National Science Foundation under Grant 2017289.

References

[1] L. Morra, S. P. Mohanty and F. Lamberti, "Artificial Intelligence in Consumer Electronics," *IEEE Consumer Electronics Magazine*, vol. 9, no. 3, pp. 46-47, 2020.

[2] J. Lemley, A. Kar, A. Drimbarean and P. Corcoran, "Convolutional neural network implementation for eye-gaze estimation on low-quality consumer imaging systems," *IEEE Trans. Consumer Electronics*, vol. 65, no. 2, pp. 179-187, 2019.

[3] J. Ding and Y. Wang, "A WiFi-based smart home fall detection system using recurrent neural network," *IEEE Trans. Consumer Electronics*, vol. 66, no. 4, pp. 308-317, 2020.

[4] R. K. Nath, H. Thapliyal, A. Caban-Holt and S. P. Mohanty, "Machine Learning Based Solutions for Real-Time Stress Monitoring," *IEEE Consumer Electronics Magazine*, vol. 9, no. 5, pp. 34-41, 2020.

[5] L. Rachakonda, S. P. Mohanty and E. Kougianos, "iLog: An Intelligent Device for Automatic Food Intake Monitoring and Stress Detection in the IoMT," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 2, pp. 115-124, 2020.

[6] M. M. Malik, “A Hierarchy of Limitations in Machine Learning”, *arXiv preprint*, arXiv:2002.05193. 2020.

[7] A. Seifert and S. Rasp, “Potential and Limitations of Machine Learning for Modeling Warm-Rain Cloud Microphysical Processes”, *Journal of Advances in Modeling Earth Systems*, 2020 Dec, 12(12), e2020MS002301.

[8] M. E. Morocho-Cayamcela, H. Lee, W. Lim, “Machine learning for 5G/B5G mobile and

wireless communications: Potential, limitations, and future directions', *IEEE Access*, 2019 Sep 19;7:137, pp. 184-206.

[9] A. Torralba and A. A. Efros, "Unbiased look at dataset bias", in *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 1521-1528, 2011.

[10] R. Geirhos, J. H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F.A. Wichmann, "Shortcut learning in deep neural networks", *Nat Mach Intell*, 2, 665–667, 2020.

[11] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints", *arXiv preprint*, arXiv:1707.09457, 2017.

[12] A. S. Rich and T. M. Gureckis, "Lessons for artificial intelligence from the study of natural stupidity", *Nature Machine Intelligence*, 1, 174, 2019.

[13] T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization", *arXiv preprint*, arXiv:1806.08010, 2018.

[14] J. Buolamwini and T. Gebru, "Gender Shades: intersectional accuracy disparities in commercial gender classification", in *Proceedings of Machine Learning Research*, 81, pp. 1–15, 2018

[15] T. Alladi, V. Chamola, B. Sikdar and K. R. Choo, "Consumer IoT: security vulnerability case studies and solutions," *IEEE Consumer Electronics Magazine*, vol. 9, no. 2, pp. 17-25, 2020.

[16] W. Z. Khan, M. Y. Aalsalem and M. K. Khan, "Communal acts of IoT consumers: a potential threat to security and privacy," *IEEE Trans. Consumer Electronics*, vol. 65, no. 1, pp. 64-72, 2019.

[17] M. Weiss, "Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions," *Technology Science*, 2019121801. December 17, 2019.

[18] DeepFaceLab. [online]. Available: https://github.com/iperov/DeepFaceLab, Last accessed on 20 April 2021.

[19] G. Lawton, "The grand delusion: Blind to bias," *New Scientist*, 2812, May 2011. Available: https://www.newscientist.com/article/mg21028122-200-the-grand-delusion-blind-to-bias/, Last accessed on 20 April 2021.

[20] T. Bolukbasi, K. -W. Chang, J. Zou, V. Saligrama, and, A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Proc. 30th International Conference on Neural Information Processing Systems*, pp. 4356–4364, 2016.

[21] R. J. Mooney, "Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning", *arXiv preprint*, cmp-lg/9612001, Dec 9, 1996.

[22] D. Brain and G.I. Webb, "On the effect of data set size on bias and variance in classification learning", in *Proceedings of the Fourth Australian Knowledge Acquisition Workshop*, pp. 117-128, 1999.

[23] K. Huang, H. Yang, I. King, M.R. and Lyu, "Learning classifiers from imbalanced data based on biased minimax probability machine", in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[24] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning", in *Proc. International Conference on Artificial Intelligence and Statistics*, pp. 702-712, 2020.

[25] L. Sweeney, "Discrimination in online ad delivery," *Communications of the ACM*, vol. 56, no. 5, pp. 44-54, May 2013.

[26] J. Angwin, N. Scheiber, and A. Tobin, "Machine bias: Dozens of companies are using Facebook to exclude older workers from job ads," *ProPublica*, Dec. 20, 2017. [Online]. Available: https://www.propublica.org/article/facebook-ads-age-discrimination-targeting

[27] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings: a tale of opacity, choice, and discrimination," in *Proc. on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 92–112, 2015.

[28] A. Pandey and A. Caliskan, "Iterative effect-size bias in ridehailing: measuring social bias in

dynamic pricing of 100 million rides", 2020. [Online]. Available: https://arxiv.org/abs/2006.04599

[29] D. Ingold and S. Soper, "Amazon doesn't consider the race of its customers: should it?" Bloomberg, April 21, 2016. [Online]. Available: https://www.bloomberg.com/graphics/2016-amazon-same-day/

[30] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, October 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[31] D. Gershgorn, "Companies are on the hook if their hiring algorithms are biased," Quartz, October 2018. [Online]. Available: https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased.

[32] J. Walker, "Meet the new boss: big data," *Wall Street Journal*, 2012. [Online]. Available: https://www.wsj.com/articles/SB10000872396390443890304578006252019616768

[33] R. Parasuraman and V. Riley, "Humans and automation: use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230-253, 1997.

[34] S. Trewin, "AI fairness for people with disabilities: point of view," *IBM Accessibility Research*, [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1811/1811.10670.pdf

[35] S. Park, J. Byun, H. Rim, D. Lee, and H. Lim, "Natural language-based user interface for mobile devices with limited resources," *IEEE Trans. Consumer Electronics*, vol. 56, no. 4, pp. 2086-2092, 2010.

[36] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases", *Science*, vol. 356, no. 6334, pp. 183-186, 2017.

[37] S. L. Blodgett, J. Wei, and B. O'Connor, "Twitter universal dependency parsing for African-American and mainstream American English," in *Proc. 56th Annual Meeting of the Association for Comput Linguistics*, pp. 1415–1425, July 2018.

[38] M. Alzantot, Y. Sharma, A. Elgohary, B. -J. Ho, M. Srivastava, and K.-W. Chang, "Generating natural language adversarial examples," in Proc. *2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2890–2896, 2018.

[39] N. Carlini and D. Wagner, Audio Adversarial Examples: Targeted Attacks on Speech-to-Text, *IEEE Security and Privacy Workshops*, pp. 1-7, 2018.

[40] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, S. Goel, "Racial disparities in automated speech recognition," in *Proceedings of the National Academy of Sciences*, vol. 117, no. 14 pp. 7684-7689 Apr 2020.

[41] D. Victor, "Microsoft created a Twitter bot to learn from users. It quickly became a racist jerk," The New York Times, March 24, 2016. [Online]. Available: https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html

[42] E. Siegel, "When does predictive technology become unethical?" *Harvard Business Review*, October 23, 2020. [Online]. Available: https://hbr.org/2020/10/when-does-predictive-technology-become-unethical

[43] R. Courtland, "Bias detectives: the researchers striving to make algorithms fair," *Nature*, vol. 558, pp. 357-360, 2018.

[44] A. Klein, "Reducing bias in AI-based financial services," Brookings AI Governance, July 10, 2020. [Online]. Available: https://www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/

[45] L. Rice and D. Swesnik, "Discriminatory effects of credit scoring on communities of color," *Suffolk University Law Review*, vol. XLVI, pp. 935-965, 2013.

[46] B. Kabler, “Insurance-based credit scores: impact on minority and low income populations in Missouri,” State of Missouri Department of Insurance, 2004. [Online]. Available: https://insurance.mo.gov/reports/credscore.pdf

[47] “Use of Credit Information by Insurers in Texas,” Texas Department of Insurance, December 30, 2004. [Online]. Available: https://www.tdi.texas.gov/reports/documents/creditrpt04.pdf

[48] C. Lecher, “Automated background checks are deciding who’s fit for a home,” The Verge. Feb 2019. [Online]. Available: https://www.theverge.com/2019/2/1/18205174/automation-background-check-criminal-records-corelogic

[49] J. Rosenberg, “DHS rule change threatens disabled care,” Arkansas Times, Oct. 2017. [Online]. Available: https://arktimes.com/news/arkansas-reporter/2017/10/12/dhs-rule-change-threatens-disabled-care

[50] D. K. Citron, “Technological due process,” U of Maryland Legal Studies Research Paper No. 2007-26, Washington University Law Review, Vol. 85, pp. 1249-1313, 2007.

[51] K. Crawford and R. Calo, “There is a blind spot in AI research,” *Nature*, vol. 538, pp. 311-313, October 2016.

[52] C. M. Gijsberts, K.A. Groenewegen, I.E. Hoefer, M.J.C. Eijkemans, F.W. Asselbergs FW, T.J. Anderson *et al.*, “Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events”, *PLoS ONE*, vol. 10, no. 7, e0132321, 2015.

[53] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, “A Machine Learning based Approach for DeepFake Detection in Social Media through Key Video Frame Extraction”, *Springer Nature Computer Science*, Vol. 2, No. 2, Article:99, 18-pages, 2021.

[54] S. Suwajanakorn, S. M. Seitz, and Ira Kemelmacher-Shlizerman, “Synthesizing Obama: learning lip sync from audio,” *ACM Trans. Graph.* vol. 36, no. 4, Article 95, 13 pages, July 2017.

[55] A. MacDonald, “YouTube admits 'wrong call' over deletion of Syrian war crime videos,” Middle East Eye, December 2017. [Online]. Available: https://www.middleeasteye.net/news/youtube-admits-wrong-call-over-deletion-syrian-war-crime-videos

[56] G. Chaslot, “YouTube’s A.I. was divisive in the US presidential election,” The Graph, Nov 2016. [Online]. Available: https://medium.com/the-graph/youtubes-ai-is-neutral-towards-clicks-but-is-biased-towards-people-and-ideas-3a2f643dea9a

[57] A. C. M. Fong and M. Usman, “Guest Editor’s Introduction: Special Section on Machine Learning for End Consumers,” *IEEE Consumer Electronics Magazine*, vol. 9, no. 5, pp. 77-78, 2020.

[58] B. Friedman and H. Nissenbaum, “Bias in computer systems,” *ACM Trans. Inf. Syst.*, vol. 14, no. 3, pp. 330–347, 1996.

[59] Data for Democracy. Available: https://datafordemocracy.org/, Last accessed on 20 April 2021.

[60] AI Index 2018 Report. Available: http://cdn.aiindex.org/2018/AI%20Index%2018%20Annual%20Report.pdf[1]https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf, Last accessed on 20 April 2021.

[61] Datasheets for Datasets. Available: https://arxiv.org/pdf/1803.09010.pdf, Last accessed on 20 April 2021.

[62] A. Kaushal, R. Altman, and C. Langlotz, “Health Care AI Systems Are Biased”, Scientific American, Nov 17, 2020. Available: https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/

[63] W. Knight, “AI Is Biased. Here's How Scientists Are Trying to Fix It”, Wired, Dec 19, 2019, Available: https://www.wired.com/story/ai-biased-how-scientists-trying-fix/

[64] K. Hao, “This is how AI bias really happens—and why it’s so hard to fix”, *MIT Technology Review*, Feb 4, 2019, Available: https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

[65] S. Saeedi, A. Panahi, S. Saeedi, and A. C. M. Fong, “CS-NLP Team at SemEval-2020 Task 4: Evaluation of state-of-the-art NLP deep learning architectures on commonsense reasoning task,” in *Proc. International Workshop on Semantic Evaluation* (SemEval-2020), 2020.

[66] G. Hong and A. C. M. Fong, “Multi-prong framework toward quality-assured AI decision making,” in *Proc. 4th International Conference on Contemporary Computing and Informatics*, pp. 106-110, 2019.

[67] The Neuro-Symbolic AI initiative. [Online]. Available: https://researcher.watson.ibm.com/researcher/view_group.php?id=10518, Last accessed on 20 April 2021.

[68] J. Silberg and J. Manyika, “Tackling bias in artificial intelligence (and in humans),” McKinsey Global Institute, June 6 2019. Available: https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

[69] A. D’Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern, “Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies”, in *Proc. Conference on Fairness, Accountability, and Transparency*, 10 pages, 2020.

[70] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva, Causal reasoning for algorithmic fairness, *arXiv preprint*, arXiv:1805.05859, 2018.

[71] A. Kasirzadeh and A. Smart, The use and misuse of counterfactuals in ethical machine learning", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[72] M. J. Kusner, J. R. Loftus, C. Russell, R. Silva, Counterfactual Fairness, *arXiv preprint*, arXiv: 1703.06856, 2018.

[73] S. Hajian, F. Bonchi, and C. Castillo, “Algorithmic bias: From discrimination discovery to fairness-aware data mining”, in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 2125-2126, 2016.

[74] A. Amini, A.P. Soleimany, W. Schwarting, S.N. Bhatia, and D. Rus, “Uncovering and mitigating algorithmic bias through learned latent structure”, in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 289-295, 2019.

[75] S. K. Ram, B. B. Das, K. Mahapatra, S. P. Mohanty and U. Choppali, "Energy Perspectives in IoT Driven Smart Villages and Smart Cities," *IEEE Consumer Electronics Magazine*, vol. 10, no. 3, pp. 19-28, May 2021.

Sirwe Saeedi is a research assistant with the Department of Computer Science at Western Michigan University. Contact her at sirwe.saeedi@wmich.edu.

Alvis C. M. Fong is a faculty member with the Department of Computer Science at Western Michigan University. Contact him at alvis.fong@wmich.edu.

Saraju P. Mohanty is currently the Editor-in-Chief for the IEEE CONSUMER ELECTRONICS MAGAZINE and a Professor with the Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA. Contact him at smohanty@ieee.org.

Ajay K. Gupta is a Professor with the Department of Computer Science at Western Michigan University. Contact him at ajay.gupta@wmich.edu.

Steve M. Carr is a Professor and Chair of the Department of Computer Science at Western Michigan University. Contact him at steve.carr@wmich.edu.