The Dangers of Drowsiness Detection: Differential Performance, Downstream Impact, and Misuses

Jakub Grzelak
jakub.grzelak@kcl.ac.uk
King's College London
London, UK

Martim Brandao
martim.brandao@kcl.ac.uk
King's College London
London, UK

ABSTRACT

Drowsiness and fatigue are important factors in driving safety and work performance. This has motivated academic research into detecting drowsiness, and sparked interest in the deployment of related products in the insurance and work-productivity sectors. In this paper we elaborate on the potential dangers of using such algorithms. We first report on an audit of performance bias across subject gender and ethnicity, identifying which groups would be disparately harmed by the deployment of a state-of-the-art drowsiness detection algorithm. We discuss some of the sources of the bias, such as the lack of robustness of facial analysis algorithms to face occlusions, facial hair, or skin tone. We then identify potential downstream harms of this performance bias, as well as potential misuses of drowsiness detection technology—focusing on driving safety and experience, insurance cream-skimming and coverage-avoidance, worker surveillance, and job precarity.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Social and professional topics** → **Surveillance**; *Race and ethnicity*; *Gender*.

KEYWORDS

drowsiness detection, bias, disparate impact, fairness, technology misuses, surveillance

ACM Reference Format:
Jakub Grzelak and Martim Brandao. 2021. The Dangers of Drowsiness Detection: Differential Performance, Downstream Impact, and Misuses. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics,and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3461702.3462593

1 INTRODUCTION

Statistics of the effect of driver sleepiness and fatigue in car accidents [13], as well as the surge of semi-autonomous driving, has motivated research into drowsiness detection (DD). These algorithms use sensor data to estimate the degree of alertness of drivers, and such estimates could potentially be used to provide warnings [9, 19] or apply other mechanisms to alert a driver [18], adjust autonomous driving settings, or flag the event to a "supervisor" for further action. Drowsiness and fatigue detection can also be used to monitor workers outside of a driving context, for example in office work [15], e-learning [11] or physically demanding work [17]. In this paper we investigate the potential harms of deploying such systems.

Many implementations of drowsiness detection use visual input from cameras and facial features to estimate drowsiness [9, 10]. However, given recent findings of computer vision algorithms consistently underperforming on specific social groups, typically ethnic minorities and already disadvantaged communities [2, 3, 12, 27], it is likely that such biases will take place in DD as well. One important question to ask is how the use of the technology and such disparities could impact different communities. Importantly, the existence of drowsiness estimates also opens the door to multiple uses and misuses of the data. For example, logistics and insurance companies are two sectors that have shown an interest in the technology[1,2], and thus its potential downstream impact can take financial and job precarity dimensions. Our goal in this paper is to anticipate and characterize the potential harms of deploying DD technology in the real world. Our contributions are the following:

(1) We conduct a bias audit of a DD algorithm, where we identify performance disparities across race and gender groups (Section 3);
(2) We identify sources of performance bias in DD, and potential downstream harms (Section 3.3);
(3) We identify potential misuses of DD technology, and discuss the gap between academic research and real-world use of DD (Section 4).

2 BACKGROUND

Various methods have been proposed in order to detect drowsiness in human subjects, including those using driving statistics [7, 19], EEG [5], EOG [15], IMU [17], thermal [14] or vision data [6, 8–10, 24]. Ramzan et al. [23] provide a comprehensive survey of drowsiness detection methods.

Vision-based methods for DD typically use facial features such as blinking rate [10], eye closure [6, 8], or yawning [14]. Some recent methods use deep neural networks to predict drowsiness from eye detections [10, 24] or face detections [9].

Some hand-designed features used in DD rely on skin segmentation [6], and deep-learning-based methods inclusively depend on face or facial feature detections [10, 24]. This dependence on skin and facial features opens the door to problems of bias given the high variability of skin tone and facial features across populations.

[1] https://www.seeingmachines.com/guardian/guardian/
[2] https://nationwidetrackingsystems.com.au/driver_behaviour/

AIES '21, May 19–21, 2021, VirtualEvent, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462593

Recent work has shown that gender recognition from facial image data often underperforms on minority groups such as black women [3], and similarly for facial analysis algorithms on older adults with dementia [26], or pedestrian detection algorithms on darker skin tones [27].

In similar spirit to the works above, in this paper we audit a state-of-the-art DD algorithm for differential performance across gender and ethnicity, and draw conclusions regarding the potential sources of bias and kinds of downstream impact. The work of Ngxande et al. [21] is related to ours in that it also audits DD algorithm performance. However, while we focus on evaluating bias across gender and ethnicity categories within the same dataset and identifying potential downstream harms, the former work [21] assesses performance on a new out-of-distribution dataset from a South-African context to evaluate the performance drop and deployment readiness. In comparison to [21] we not only audit an algorithm but also identify potential downstream harms and misuses of the technology.

3 DIFFERENTIAL PERFORMANCE AUDIT

3.1 Experimental setup

We audit the temporal model proposed by Ghoddoosian et al. [10] for the Real-Life Drowsiness Dataset (RLDD) [10]. The method uses hand-designed blink features as input to an LSTM that predicts drowsiness on a scale from 0 to 10 (Extremely Alert to Extremely Sleepy). We chose this method due to its use of facial features, which are a common predictor used across multiple methods [6, 8, 24], and also due to the availability of source code using the RLDD dataset—which contains information about gender and ethnicity of participants.

The RLDD dataset consists of 180 RGB videos, each approximately 10 minute long. Videos are separated into 5 folds, each fold consisting of videos from 12 participants. There are 60 participants in total (annotated as 51 Male and 9 Female) from different ethnicities (10 Caucasian, 5 Non-white Hispanic, 30 Indo-Aryan & Dravidian, 8 Middle Eastern, and 7 East Asian). These demographic labels were self-reported by participants. Labels are not made publicly available with the rest of the data, but total counts of each category are reported in the original publication [10]. Therefore, we (the authors of this paper) manually labeled each of the participants into gender and ethnicity categories after watching the participants' videos[3]. Our annotations were consistent with the label-counts above. The videos in the dataset were obtained by each participant filming themselves in three different drowsiness states: alert, low vigilant or drowsy. Drowsiness labels were also provided by participants themselves according to how they were feeling: alert, low vigilant, or drowsy.

In our experiments, we first trained the method to similar performance reported by Ghoddoosian et al. [10], and then computed performance metrics over each subset of participants (i.e. performance over Male-, Female-, Caucasian-, ...-labeled participants). We use the same performance metrics as reported in [10]: Blink Sequence Accuracy (BSA), Blink Sequence Regression Error (BSRE), Video Accuracy (VA) and Video Regression Error (VRE). Similarly to that publication we focus most comparisons on the VA metric, as it is more appropriate for evaluating performance over a whole population [10]. We computed these metrics in three different conditions, described below.

[3]We started by labeling all participants without checking the total counts from the original paper. We left participants to whom group assignment was unclear to the end. There were two of these remaining participants. Group membership of these became clear once we had access to the remaining labels.

Table 1: Performance of our obtained model compared to authors' published results. Values averaged across all test folds.

Result	BSRE	VRE	BSA	VA
[10]	1.90	1.14	54.0%	65.2%
ours	1.59	1.28	56.6%	67.2%

3.1.1 Original condition. This experiment is a reconstruction of the original results in [10]. As in the original publication, we use one fold as a test set and the remaining four folds for training; we train the model multiple times from different random seeds (10 times in our experiments); and report results on the models that perform best on the respective test sets (i.e. for each training-test split we run training 10 times on the training set to obtain 10 models, pick the model with best VA performance on the test set). These results, therefore, are in practice using the test set for training.

3.1.2 Isolated training-and-test condition. In this experiment we conduct the same procedure as in the previous condition, except we pick the model (out of 10) that performs best on the training set itself. Then, we compute performance results on the test set using the chosen model, and these are the results we report in the paper. This is to avoid using test data in training.

3.1.3 Missing group condition. In the final experiment we isolate the training and test sets as in the previous condition. However, this time we evaluate the models' capability to generalize to new groups. Specifically, for each group we compute the method's performance on that group using a model trained on all other groups (e.g. performance on Female subjects computed from a model trained only on Male subjects).

3.2 Results

3.2.1 Original condition experiment. Table 1 and 2 show the performance of the models on the full dataset, and the original results reported in [10]. The results are similar, and VA is actually 2% higher than that reported by [10]. This could be related to random factors in training, e.g. initialization of the models or random shuffle of samples each epoch.

Table 3 shows the performance measured over each subset of participants. To obtain the performance of the models on a certain group (e.g. VA on Female) we first computed each participant's prediction using the appropriate model—i.e. using the model that was not trained on that participant. Then, we computed the performance metric over all participants in that group. The table shows that the models performed worse for Indo-Aryan & Dravidian, Middle Eastern, and East Asian groups (compared to Caucasian and Non-white Hispanic), and for the Male group (compared to Female).

Table 2: Confusion matrices for the original condition experiment. Our (top) vs authors' results (bottom).

	Alert	Low vigilant	Drowsy
Alert	0.80	0.10	0.05
Low vigilant	0.13	0.42	0.15
Drowsy	0.07	0.48	0.80
	Alert	Low vigilant	Drowsy
Alert	0.81	0.12	0.05
Low vigilant	0.18	0.32	0.13
Drowsy	0.01	0.56	0.82

Table 3: Performance of the model in the original condition, for different genders and ethnicities.

Group	BSRE	VRE	BSA	VA
Male	1.59	1.40	55.4%	66.0%
Female	1.48	0.63	59.5%	74.1%
Caucasian	1.13	1.27	65.2%	76.7%
Non-white Hispanic	0.81	0.29	65.2%	73.3%
Indo-Aryan & Dravidian	1.99	1.40	51.3%	63.3%
Middle Eastern	1.97	1.60	48.1%	66.7%
East Asian	1.11	1.13	60.1%	66.7%

The VA for Male participants was lower than Female by 8.1 percentage points; and the VA differences in ethnicity ranged from 6.6 percentage points (i.e. between Middle Eastern and Non-white Hispanic) to 13.4 points (between Indo-Aryan & Dravidian and Caucasian). Both regression errors, BSRE and VRE, for Indo-Aryan & Dravidian, and Middle Eastern were also higher than for other ethnicity groups (Table 3).

Table 4 shows that on average Female and East Asian groups had a similar number of datapoints used for training, where a datapoint is a sequence of 30 consecutive blinks captured in a video. However, the method performed worse on the East Asian group (66.7% VA) than on the Female (74.1% VA), as seen in Table 3. In contrast, Indo-Aryan & Dravidian, and Middle Eastern had both a similar amount of datapoints and similar VA (66.7% and 63.3%). Finally, Non-white Hispanic subjects had the lowest average number of datapoints, 87, but high performance (73.3% VA). Performance was therefore not straightforwardly related to the number of detected blinks.

Table 5 shows the confusion matrix for the Middle Eastern group as an example. The table shows the method is biased towards a "drowsy" state, as the model wrongly predicted that state instead of "Low Vigilant" in 75% of the cases. As we will discuss later on, this observation has important consequences for downstream applications of the method, such as insurance premiums or driver-assist system performance.

3.2.2 Isolated training-and-test condition experiment. Table 6 shows that in this second experiment, where we avoided using the test set in training, Video Accuracy was 72.6% on the training set and 59.4% on the test set. This result shows that real-world performance should be considerably lower than that reported originally in [10].

Table 7 shows per-group performance on this condition. Similarly to the original-condition experiment, the method performed worse on Indo-Aryan & Dravidian, Middle Eastern, and Male groups. For example, Male VA was almost 13 percentage points lower than Female, and performance on the Middle Eastern group was 16.7 percentage points lower than Caucasian.

Confusion matrices for the Middle Eastern and Indo-Aryan & Dravidian groups in Table 8 show that the method is also biased

Table 4: Average number of contributed datapoints per subject.

Group	Avg. num. datapoints per subject
Male	113
Female	233
Caucasian	173
Non-white Hispanic	87
Indo-Aryan & Dravidian	112
Middle Eastern	107
East Asian	214

Table 5: Confusion matrix for Middle Eastern subjects in the original condition.

	Alert	Low vigilant	Drowsy
Alert	0.75	0	0
Low vigilant	0.13	0.25	0
Drowsy	0.12	0.75	1.00

Table 6: Performance of the model on the "isolated training-and-test" condition. Values are averaged over all test folds.

Group	BSRE	VRE	BSA	VA
Training set	0.85	0.74	68.9%	72.6%
Test set	1.59	1.08	55.2%	59.4%

Table 7: Performance of the model for different genders and ethnicities in the "isolated training-and-test" condition.

Group	BSRE	VRE	BSA	VA
Male	1.56	1.22	54.4%	57.5%
Female	1.40	0.30	61.0%	70.4%
Caucasian	0.88	0.39	65.4%	66.7%
Non-white Hispanic	1.30	1.58	61.8%	66.7%
Indo-Aryan & Dravidian	1.85	1.17	53.8%	56.7%
Middle Eastern	2.18	1.33	40.8%	50.0%
East Asian	1.23	1.08	57.8%	66.7%

Table 8: Confusion matrices for: Middle Eastern (top) and Indo-Aryan & Dravidian (bottom) in the "isolated training-and-test" condition.

	Alert	Low vigilant	Drowsy
Alert	0.63	0	0
Low vigilant	0.37	0.12	0.25
Drowsy	0	0.88	0.75
	Alert	Low vigilant	Drowsy
Alert	0.90	0.17	0.10
Low vigilant	0.10	0.17	0.27
Drowsy	0	0.66	0.63

Table 9: Performance of the model for different genders and ethnicities in the missing group condition.

Group	BSRE	VRE	BSA	VA
Male	2.06	1.78	51.2%	56.2%
Female	1.45	0.58	61.7%	81.5%
Caucasian	0.93	0.77	61.3%	60.0%
Non-white Hispanic	0.87	0.54	62.9%	66.7%
Indo-Aryan & Dravidian	2.08	1.96	56.7%	58.9%
Middle Eastern	1.91	1.04	47.1%	62.5%
East Asian	0.94	0.61	59.5%	47.6%

towards the "drowsy" state in this condition. For Middle Eastern, in 88% of the cases when they felt low vigilant, the model predicted that they were drowsy. For Indo-Aryan & Dravidian, this occurred in 66% of the cases. Such high bias was not present in other groups (20%, 20% and 43% for Caucasian, Non-white Hispanic, and East Asian, respectively).

3.2.3 Missing group condition experiment. In the final experiment we computed the performance on each group when the group is excluded from training. The results, shown in Table 9, are similar to the previous conditions: the method performed poorly on Male, Caucasian, Indo-Aryan & Dravidian and East Asian groups, as Video Accuracy was below 60.0% for each of these groups. Performance on Male subjects was 25.3 percentage points lower than Female. Comparing Tables 7 and 9 reveals that the method generalizes poorly to unseen classes, as the performance drops considerably on groups excluded from training. For example, VA dropped by 6.7 percentage points on the Caucasian group and by 19.1 points on the East Asian group. Video Regression Error for Indo-Aryan & Dravidian and Male groups is also the highest in this experiment.

3.3 Discussion of results

3.3.1 Consistent results. The results in all conditions showed consistent performance disparities between certain groups. The method performed worse on Male than Female subjects, and this was consistent both across metrics and conditions. This could seem surprising given that there are fewer videos labeled Female. However, the Female group actually had more datapoints to use for training (a datapoint is a sequence of 30 consecutive blinks captured in a video).

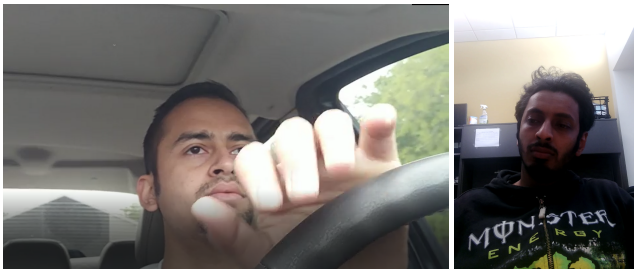

Figure 1: Examples of blink-detection failure.

The methods also performed worse consistently on Indo-Aryan & Dravidian, Middle Eastern, and East Asian groups. Performance for these groups was the lowest in all conditions. Furthermore, the models could be especially harmful for Middle Eastern and Indo-Aryan & Dravidian groups, as they were biased towards a drowsy prediction in these groups.

3.3.2 The source of disparity. The drowsiness detection algorithm considered in the experiments uses four blink features: duration, amplitude, eye-opening velocity and blinking frequency. The RLDD dataset contains 9 videos labeled Female, who on average generated more datapoints than Male. This suggests that detection performance of these four blinking features could be negatively affected by some gender-specific appearance characteristics. After inspection (by the authors of this paper) of all Male videos we identified 24 Male participants had significant facial hair, shade over the eyes, or dark skin tone. Examples of these images are shown in Figure 1. This could affect measurements of eye-opening velocity and amplitude. Moreover, Male participants often touched their faces and beards, leading to blinks being misdetected or undetected—and decreasing drowsiness detection performance.

Excluding a group from the training set in the third condition had a strong impact on the results. This suggests that there could be blinking features characteristic of specific groups. It also could be the case that physical traits of a given gender or ethnicity make it more difficult to predict drowsiness state. Finally, it is possible that there are group-consistent differences in the way participants estimate their own drowsiness states—such as a consistent under-estimation or over-estimation of their own drowsiness state on some groups.

3.3.3 Potential downstream disparate impact. Assuming drowsiness detection systems can achieve high prediction performance and we can find effective feedback mechanisms to help drivers control drowsiness, then such systems could indeed be useful tools to prevent sleep-related car accidents. However, the results of our experiments show that these methods can lead to the opposite outcome. For people of an ethnicity that the method performs poorly on, the use of a drowsiness detection system could be more dangerous than not using the system at all. For example, consider the following hypothetical systems:

System 1: If the system is such that it does not let a person fully

drive the car (e.g. make sudden turns, breaks, or accelerations) when the driver is thought not to be alert, then such functionality could prevent some people from doing proper driving. If the system used the method we audit in this paper, East Asian and Middle Eastern groups would be more subject to driving limitations even when they were awake, due to mis-classification.
System 2: If the system is such that it does not let a driver *turn on and start driving* a car, this could also lead to considerable harm—such as in emergency situations due to labor or heavy injury, where not being able to use the car would be disastrous and unfair, but also in mundane situations due to harms from travel delays.
System 3: If the system is such that it uses sounds or physical stimuli to raise the drivers' state of alert, East Asian and Middle Eastern drivers would in our case be unnecessarily stimulated, thus potentially leading to greater discomfort, stress or annoyance when compared to other groups. Such physical or mental states could actually lower drivers' attention or overall driving performance, and increase the chance of an accident.

Another example of downstream impact is on insurance premium and coverage discrimination. If an insurance company used drowsiness predictions to inform premiums and coverage, the models audited in this paper would disparately harm Middle Eastern and Indo-Aryan & Dravidian groups, for which drowsiness detection is biased towards the "drowsy" state. In these cases, an insurance company could unfairly avoid covering an accident due to a mistaken drowsiness estimate, or it could consistently raise premiums for people of these groups even if they were not less awake on average.

3.3.4 Technical solutions and barriers. The RLDD dataset appears to be diverse as it includes subjects from five different ethnicities. However, many other ethnic and gender groups could have been represented. The method audited here did not generalize well to unseen classes, and our results suggest there could be group-specific features or behavior that is predictive of drowsiness. Therefore, manufacturers of drowsiness detection algorithms should avoid using RLDD and the temporal model audited here in production.

Our analysis also showed that there are important factors correlated with performance differentials, such as facial hair, dark skin tone, and face-touching habits. Thus, a varied dataset in terms of facial features, and a rigorous analysis of both detector and model performance across groups is necessary for responsible development in this area.

Finally, there is a need for realistic driving data, where drowsiness states are not self-reported but measured through effective sensors or physical exams. Almost all videos in the RLDD dataset are recorded indoors as participants looked at a screen, but drivers will behave differently. Drivers may move their head and gaze in different ways, wear sunglasses, or touch their faces more often. All of these factors can lead a facial-feature-based drowsiness detection algorithm fail—and so care should be taken to address this point in real-world methods.

3.3.5 Experiments' limitations. The main limitation of our analysis is the fact that we audited a single drowsiness detection algorithm on a single dataset. There are other drowsiness detection methods which take into consideration different factors in their training process. For instance, Reddy et al. [24] take into consideration eyes, mouth and face crops; and [20] uses end-to-end architectures. However, our observation of a lack of generalization was also seen in the work of [20] for end-to-end architectures.

Another limitation of our work is the categorization used for gender and ethnicity. Even though categories were self-reported, this data is not provided as part of the dataset and we thus had to label each participant's group membership. Although our group counts matched those reported in [10], potential errors in group labeling could still exist and thus introduce errors in the analysis. The lack of representation of other groups, and the explicit use of discrete categories itself could also perpetuate wrong conceptions of gender and race. We believe the results are still in the interest of minorities and marginalized groups as they show potential sources of bias and downstream harm.

4 POTENTIAL USES AND MISUSES

4.1 Promoted applications and uses in research

Drowsiness detection is often promoted in research papers as a tool to increase safety and self-awareness. DD has been applied to driver monitoring [9, 19], e-learning [11], air traffic control worker monitoring [25], industry worker monitoring for productivity and accident prevention [17], and office worker monitoring [15].

In a Scopus[4] search of all papers mentioning the words "drowsiness detection", "sleepiness detection" or "fatigue detection" in their title or abstract, we identified 1663 papers, out of which driver monitoring was the most popular application (73% papers), followed by worker monitoring (2%) and education (1%)[5].

Research papers often state that drowsiness detection can be used to decide when to hand over control to an autonomous (driving) system [24]; use warnings [9, 19] or music change [18] to increase driver alert state; suggest driving/working breaks [15], or simply inform users of their own state [17]. Outside of the driving context, alert levels are assumed to be useful for evaluating and iterating user interface design [1] and to adapt lesson plans in e-learning scenarios [11]. The overarching use of DD as promoted in research is thus to increase human safety and efficiency through surveillance. Correspondingly, 50% of the 1663 Scopus papers mention "safety" or "accidents" in the abstract, while 14% mention "efficient" or "efficiency".

One criticism that can be made of this narrative of DD *for safety and efficiency* is that it misses the fact that the most interested parties in using such technology are likely those that have financial gain in using it—such as health and car insurance, and companies looking to enforce worker productivity in financial terms—thus leading to issues of discrimination and exploitation as we will see next.

Another gap in the academic literature is that of a lack of short- and long-term evaluation of the full (feedback) system. In an attempt to decompose the problem, researchers assume that downstream mechanisms will be able to solve the consequences of a drowsiness problem, and researchers thus focus on the actual performance of the detection algorithm in isolation. However, performance of the full system is crucial for the understanding and proper regulation of the technology.

[4] https://www.scopus.com/
[5] driving application identified by keywords: "driver" or "driving", education: "teaching" or "education", worker monitoring: "worker" or "employee" keywords.

4.2 Potential misuses of DD technology

We identify two potential misuses of DD: insurance misuse and worker surveillance.

4.2.1 Insurance misuse. In the case of driver monitoring applications, drowsiness data is appealing to insurance companies as one more predictor of risk, and to adjust premiums. Even though "insurance" is not mentioned in research papers as a potential use for drowsiness detection (0 entries in the 1663 Scopus abstracts), such use is most realistic in the context of driver monitoring—as there are clear financial gains for insurance companies. Fine-grained risk measurements allow insurers to refine risk categories, allowing lower-risk groups to pay lower premiums at the cost of higher premiums on high-risk groups. However, such refinement leads to what is known as "cream skimming" [4], where insurers end up excluding high-risk groups from the service—and only provide insurance to those that are less likely to need it. Insurance companies also have an interest in using drowsiness detection as a device to blame a driver and avoid responsibility in accident coverage.

Drowsiness and fatigue detection is already provided as a product by some companies[6], which advertise it as "life-saving" technology that prevents accidents through warning interventions. It is targeted at logistics companies that want to protect their drivers' safety, but also to "determine if they are distracted while driving". In fact, some insurance companies already use this product in their services in order to "help [customers] save on vehicle insurance" [7].

Such discourse raises a second question about drowsiness detection, which is how this data will be used by insurance providers and employers once it is made available. For example, logistics companies may use detection data (regardless of its accuracy) as a surveillance device to challenge drivers' self-reported fatigue-levels or justify driver termination, regardless of both the algorithm's accuracy and the source of drowsiness—which could come from companies' bad scheduling and work-time practices.

4.2.2 Worker surveillance. When DD is used for worker surveillance by employers and managers, the promise is once more that of safety. However, issues of worker control and organizational malpractices may ensue from the availability of drowsiness detection data in such contexts. For example, the truck industry in the US (which consists of roughly 3 million drivers [22]) already uses electronic monitoring devices intensively to manage work hours, inspections, breaks, delivery speeds, etc. As shown by Levy [16], such devices also produce considerable data related to driving efficiency, idling time, acceleration patterns and other factors—which managers then use as leverage to convince drivers to work longer hours. One common practice is to use live tracking of working hours and a comparison to other drivers in order to challenge a driver's claim of high fatigue (and force them shorten/postpone a break) [16]. The industry's eagerness to use such devices for control, and current malpractices based on indirectly estimating fatigue, indicates a potential interest in DD technology. If drowsiness-estimation data also became available to managers and dispatchers, such data could be used to further increase authority over drivers and the ability to challenge drivers' accounts—e.g. "you still have working hours left and the system says you are not tired, so you have to skip this break". However, as shown by Levy [16], drivers feel that they always know their biophysical condition best, and are hurt personally and professionally when told by others about their condition. Another practice that is used in the truck industry also sheds some light into the ways in which drowsiness detection algorithms could end up being used. Multiple companies explicitly compare drivers against each other through public leader-boards of a performance metric (e.g. per-driver fuel or time efficiency) and thus create social pressure for drivers to do longer hours or faster driving [16]. If similar practices were applied by management to signals of drowsiness-estimates, through drowsiness-per-mile or similar rankings, they could introduce new dimensions of control and unhealthy work incentives—as well as potentially serving as excuses to fire less "fit" drivers. Finally, such worker control issues would be further complicated by the fact that drowsiness detection estimates can be specifically poor on the Male group (Section 3)—which constitutes 92% of the trucker workforce [22].

[6] https://www.seeingmachines.com/guardian/guardian/
[7] https://nationwidetrackingsystems.com.au/driver_behaviour/

5 CONCLUSION

In this paper we analyzed the performance of a state-of-the-art drowsiness detection algorithm across subjects of different gender and ethnicity. We showed that the algorithm performs consistently worse on Male compared to Female participants, as well as on Indo-Aryan & Dravidian, Middle Eastern, and East Asian groups compared to Caucasian and Non-white Hispanic. We showed that the algorithm does not generalize well to unseen groups, and that there are thus potentially group-specific features for drowsiness detection (or their subjective self-reported values). We identified potential sources of this performance differential, such as facial hair, skin tone, and face-touching frequency.

We also identified potential downstream harms of this bias, in terms of disparate impact in safety, user annoyance, insurance premiums and coverage, and work precarity. We ended up with a discussion of potential misuses of the technology. While DD is currently promoted in research as a device for safety and efficiency, it is in practice a door into worker surveillance and precarity, and insurance coverage and premium discrimination.

This paper thus serves as a warning to the kinds of impact and misuses of drowsiness detection technology, and as a resource for reflection and policy regarding the use of this technology. It also highlights the gap between academic research on DD—specifically the narratives of safety and efficiency it promotes—and the potential real-world interests and risks. These are, with few exceptions [20, 21], left out of academic discussion, but need further exploration and consideration.

REFERENCES

[1] Evgeniy Abdulin and Oleg Komogortsev. 2015. User eye fatigue detection via eye movement behavior. In *Proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems*. 1265–1270.
[2] Martim Brandao. 2019. Age and gender bias in pedestrian detection algorithms. In *Workshop on Fairness Accountability Transparency and Ethics in Computer Vision, CVPR*.
[3] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine*

Learning Research, Vol. 81), Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91.

[4] David A Cather. 2018. Cream skimming: innovations in insurance risk classification and adverse selection. *Risk Management and Insurance Review* 21, 2 (2018), 335–366.

[5] Agustina Garcés Correa, Lorena Orosco, and Eric Laciar. 2014. Automatic detection of drowsiness in EEG records based on multimodal analysis. *Medical engineering & physics* 36, 2 (2014), 244–249.

[6] Bogusław Cyganek and Sławomir Gruszczyński. 2014. Hybrid computer vision system for drivers' eye recognition and fatigue monitoring. *Neurocomputing* 126 (2014), 78–94.

[7] Pia M Forsman, Bryan J Vila, Robert A Short, Christopher G Mott, and Hans PA Van Dongen. 2013. Efficient driver drowsiness detection at moderate levels of drowsiness. *Accident Analysis & Prevention* 50 (2013), 341–350.

[8] I Garcia, Sebastian Bronte, Luis Miguel Bergasa, Javier Almazán, and J Yebes. 2012. Vision-based drowsiness detector for real driving conditions. In *2012 IEEE Intelligent Vehicles Symposium*. IEEE, 618–623.

[9] Miguel García-García, Alice Caplier, and Michèle Rombaut. 2018. Sleep deprivation detection for real-time driver monitoring using deep learning. In *International Conference Image Analysis and Recognition*. Springer, 435–442.

[10] Reza Ghoddoosian, Marnim Galib, and Vassilis Athitsos. 2019. A realistic dataset and baseline temporal model for early drowsiness detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[11] SL Happy, Anirban Dasgupta, Priyadarshi Patnaik, and Aurobinda Routray. 2013. Automated alertness and emotion detection for empathic feedback during e-Learning. In *2013 IEEE Fifth International Conference on Technology for Education (t4e 2013)*. IEEE, 47–50.

[12] Ayanna Howard, Cha Zhang, and Eric Horvitz. 2017. Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*. IEEE, 1–7.

[13] Sheila G Klauer, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, David J Ramsey, et al. 2006. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. (2006).

[14] Mateusz Knapik and Bogusław Cyganek. 2019. DriverâĂŹs fatigue recognition based on yawn detection in thermal images. *Neurocomputing* 338 (2019), 274–292.

[15] Marcin Kołodziej, Paweł Tarnowski, Dariusz J Sawicki, Andrzej Majkowski, Ramigiusz J Rak, Aleksandra Bala, and Agnieszka Pluta. 2020. Fatigue Detection Caused by Office Work With the Use of EOG Signal. *IEEE Sensors Journal* 20, 24 (2020), 15213–15223.

[16] Karen EC Levy. 2015. The contexts of control: Information, power, and truck-driving work. *The Information Society* 31, 2 (2015), 160–174.

[17] Ping Li, Ramy Meziane, Martin J-D Otis, Hassan Ezzaidi, and Philippe Cardou. 2014. A Smart Safety Helmet using IMU and EEG sensors for worker fatigue detection. In *2014 IEEE International Symposium on Robotic and Sensors Environments (ROSE) Proceedings*. IEEE, 55–60.

[18] Ning-Han Liu, Cheng-Yu Chiang, and Hsiang-Ming Hsu. 2013. Improving driver alertness through music selection using a mobile EEG to detect brainwaves. *Sensors* 13, 7 (2013), 8199–8221.

[19] Anthony D McDonald, John D Lee, Chris Schwarz, and Timothy L Brown. 2018. A contextual and temporal algorithm for driver drowsiness detection. *Accident Analysis & Prevention* 113 (2018), 25–37.

[20] Mkhuseli Ngxande, Jules-Raymond Tapamo, and Michael Burke. 2020. Bias Remediation in Driver Drowsiness Detection Systems Using Generative Adversarial Networks. *IEEE Access* 8 (2020), 55592–55601.

[21] Mkhuseli Ngxande, Jules-Raymond Tapamo, and Michael Burke. 2020. Detecting inter-sectional accuracy differences in driver drowsiness detection algorithms. In *2020 International SAUPEC/RobMech/PRASA Conference*. IEEE, 1–6.

[22] Bureau of Labor Statistics. 2020. Labor force statistics from the current population survey. https://www.bls.gov/cps/cpsaat11.pdf Accessed on 2021-04-19.

[23] Muhammad Ramzan, Hikmat Ullah Khan, Shahid Mahmood Awan, Amina Ismail, Mahwish Ilyas, and Ahsan Mahmood. 2019. A survey on state-of-the-art drowsiness detection techniques. *IEEE Access* 7 (2019), 61904–61919.

[24] Bhargava Reddy, Ye-Hoon Kim, Sojung Yun, Chanwon Seo, and Junik Jang. 2017. Real-time driver drowsiness detection for embedded system using model compression of deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 121–128.

[25] Zhiyuan Shen, Guozhuang Pan, and Yonggang Yan. 2020. A High-Precision Fatigue Detecting Method for Air Traffic Controllers Based on Revised Fractal Dimension Feature. *Mathematical Problems in Engineering* 2020 (2020).

[26] B. Taati, S. Zhao, A. B. Ashraf, A. Asgarian, M. E. Browne, K. M. Prkachin, A. Mihailidis, and T. Hadjistavropoulos. 2019. Algorithmic Bias in Clinical PopulationsâĂŤEvaluating and Improving Facial Analysis Technology in Older Adults With Dementia. *IEEE Access* 7 (2019), 25527–25534. https://doi.org/10.1109/ACCESS.2019.2900022

[27] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive Inequity in Object Detection. *arXiv preprint arXiv:1902.11097* (2019).