Inclusive Ethical AI In Human Computer Interaction in Autonomous Vehicles

Sudha Jamthe
Stanford University CSP
sujamthe@businessschoolofai.com

Suresh Lokiah
Independent Researcher
Sureshlokiah@gmail.com

Yashaswini Viswanath
RACE, Reva University
yashaswini.cse@gmail.com

Abstract

Artificial Intelligence (AI) used in Autonomous Vehicles (AV) to check for driver alertness is a critical piece of technology that makes the decision to hand over control to the human if there is a disengagement of the autonomous capability. It is important that this AI be inclusive without bias because treating drivers differently will impact the safety of humans not only in the vehicle but also on the road. This paper evaluates the AI that powers driver attention systems in the car to check if the AI treats all humans inclusively the same way beyond their ethnicity, gender and age and whether it follows AI Ethics principles of Trust, Transparency and Fairness. Driver attention is built using two different AI models. One uses camera data to recognize humans and the other evaluates whether the human is alert. We found that both these AI models are biased and not inclusive of all people in all situations. We also found that there are unethical practices in how humans are tracked to check for alertness by using infrared sensors that track their retina movements without any concept of consent or privacy of people being tracked in the vehicles. This paper builds upon prior research on face detection outside the car[1] and research that shows that Car Cognition AI does not recognize all humans on the road equally[2]. We present research results about how the car is biased against some humans in its face identification and how the assertion of alertness of humans to handover control during an emergency is fundamentally

[1] Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558).

[2] Wilson, Benjamin, Judy Hoffman, and Jamie Morgenstern. "Predictive inequity in object detection." arXiv preprint arXiv:1902.11097 (2019).

flawed in its definition of alertness. We recommend mitigation techniques and call for further research to build upon our work to make the autonomous vehicle inclusive with bias mitigation of bias in all forms of AI in autonomous vehicles.

Keywords: autonomous cars, self-driving cars, driverless cars, bias, technology, robotaxi, driverless world, full autonomy, autonomous vehicles, ADAS, connected vehicles, data in the car, AI in the car, data, artificial intelligence, AI Ethics , Inclusive AI.

Introduction

AI powers Autonomous Vehicles (AV) by giving them perception to see the road. Previous research papers have covered AI Ethics in AV from the context of the trolley problem[3] and what should be coded in how the car should decide how many lives to save in a collision. The reality today is that the AV runs in Level 3 or Level 4 autonomy which defines their levels of autonomy[4] where the AI will hand over control to a human when they cannot manage running autonomously in some situation or road conditions. So, it is more common for an AV to lose control and look for a human to hand over control than a hypothetical situation where the AV has to make decisions between two sets of humans on the road. In rare situations, the human might be remotely monitoring the vehicle. In most cases, the AV will plan to give control to the human in the vehicle when such a disengagement occurs. This creates an urgent need for us to look at the AI in the car that makes judgements and predictions about people inside the vehicle without their consent and monitors them and decides if they are alert or awake. This paper focuses on AI in the vehicle that powers driver attention systems and we tested to see if this AI is inclusive AI, in that it supports all humans regardless of race, gender or other factors. To evaluate driver attention systems, first we need to understand AI in the car, what does Inclusive Ethical AI mean and what is the role of humans inside the vehicle and their relationship to the AI.

I. What is AI in the Car?

Autonomous Vehicles are built using a variety of AI technologies. The vehicle is

[3] Nyholm, Sven, and Jilles Smids. "The ethics of accident-algorithms for self-driving cars: An applied trolley problem?." *Ethical theory and moral practice* 19.5 (2016): 1275-1289.

[4] SAE International Releases Updated Visual Chart for Its “Levels of Driving Automation” Standard for Self-Driving Vehicles, 2018-12-11 WARRENDALE, PA. https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles

fitted with sensors, radars, Lidars and cameras to watch its surroundings. Sensor fusion is the technique to combine the data from these sensors to help the vehicle get a mental map of its surroundings. This map shows the road, lanes, sidewalks, shrubbery, other cars, people and traffic lights as objects that are identified by the car cognition AI. The AI collects this information dynamically as vectors to show what is static and the direction of movement of other cars and people to create perception for the car to do path planning to decide to stop, move or make a turn. This is the foundation of Autonomous vehicle technology that constantly keeps looking at humans on the road to avoid collisions.[5]

Inside the vehicle there is AI that facilities interaction with the human in the car. This includes voice assistants, augmented reality, and infotainment and some invisible AI from the data in the vehicle. This AI helps make decisions about cybersecurity, and driver assistance and to predict driving behavior to ensure safety of the people in the vehicle. All such AI interact with humans in the vehicles without any transparency on how inclusive they are in recognizing all people in all situations. The most critical of these is the AI that decides whether to give control to a human in the vehicle when the AI fails and needs the human to stop a collision.

ADAS capabilities in OEM cars

Advanced Driver Assistance Systems (ADAS) are a set of technologies incorporated in cars by OEMs to improve road safety to avoid accidents caused by humans. These technologies include sensors and camera data to help the driver drive safely and can be incorporated as an alert to the driver or in some instances automatically maneuver the vehicle to avoid collision. Some examples of ADAS features are listed in Table 2. An example of an ADAS feature can be as simple as automatically turning on the car lights when it gets dark outside to warn the driver about possible obstacles to avoid on the road.

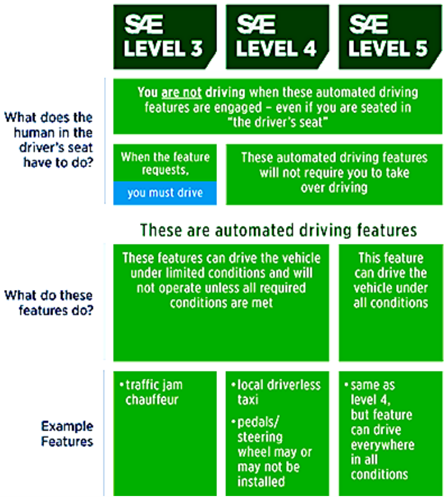
OEM car manufacturers are adding autonomous AI capabilities as ADAS features to the existing cars to enhance the security of human driver cars. For example “Forward Collision" ADAS feature uses proximity sensors to check if the car is getting too close to the vehicle in front of it. The AI then decides to alert the driver or turn on adaptive cruise control to automatically reduce the speed of the vehicle to avoid collision. There is no transparency about the AI that powers these decisions inside the vehicle.

Levels of Autonomy in Autonomous Vehicles

Autonomy in self-driving cars is measured as Level 1 to Level 5 using SAE’s Autonomy levels. Standards from Society of Automotive Engineering (SAE) International has published “SAE J3016” in 2015 and revised it in 2018 to describe 5 levels of vehicle automation from no automation to full automation. This is the de facto industry standard following the adoption by the US Department of Transportation.

[5] “2030 The Driverless World: Business Transformation from Autonomous Vehicles”, Sudha Jamthe, Sep 2017 (https://www.amazon.com/gp/product/1973753677/)”

- Level 5 is full autonomy where the vehicle can drive autonomously without human support in all road conditions.
- Level 4 is the ability to drive autonomously but within a specific set of roads or weather conditions for which the vehicle is trained.
- OEMs offer autonomous features as Level 2 or Level 3 Advanced Driver Assistance System (ADAS) features. Level 3 is where a human driver is required to be on standby to take control when needed.



	SAE LEVEL 3	SAE LEVEL 4	SAE LEVEL 5
What does the human in the driver's seat have to do?	You are not driving when these automated driving features are engaged – even if you are seated in "the driver's seat"		
	When the feature requests, you must drive	These automated driving features will not require you to take over driving	
	These are automated driving features		
What do these features do?	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met		This feature can drive the vehicle under all conditions
Example Features	• traffic jam chauffeur	• local driverless taxi • pedals/ steering wheel may or may not be installed	• same as level 4, but feature can drive everywhere in all conditions

Image1: SAE J3016 Levels of Autonomy

II. Role of humans in-car and agency of the vehicle

AI today has the agency to decide when it is not confident to drive autonomously any more. The AI detects and judges any human inside the vehicle and hands over control of the vehicle. The ADAS features also incorporate each autonomous feature differently based on whether the vehicle is operating at level 2 or level 3 autonomy. In Level 2 Autonomous Vehicle the AI informs the user of a “Forward Collision Warning" or a “Lane Departure Warning.” In Level 3 Autonomous Vehicle, the AI will decide to take charge and reduce the speed of the vehicle to avoid a collision or steer the car back into the lab if it detects that the vehicle is not staying within a lane.

The human can be a standby driver or a passenger inside the vehicle and has no agency in the decision of the safety situations of the AI even when the AI loses control of the vehicle.

This AI is the focus of this paper on whether it is an ethical and inclusive AI or not and what are the repercussions of any bias of this AI against any segment of the population.

Reference: Table 1: List of Advanced Driver Assistance System (ADAS) features

Advanced Driver Assistance System (ADAS) features	Autonomous feature explained
1. Lane departure warning	AI checked in the vehicle stays within lane markings on the road
2. Traffic signal recognition	AI looks for traffic lights to get car ready to stop
3. Forward collision warning	AI looks if vehicle is too close to car in the front to reduce speed to prevent a collision
4. Driver Alertness Detection	AI detects if the driver is alert and warns if they are sleepy or decides if they are alert to give control in the case an autonomous vehicle disengagement

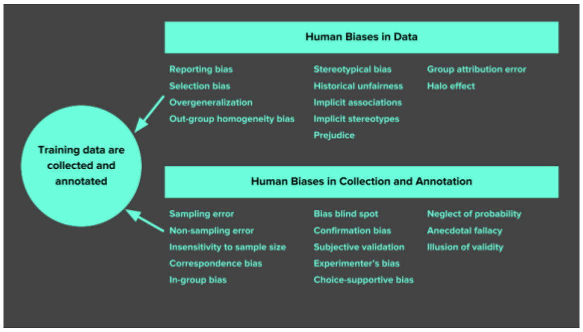
III. What are AI Ethics and Inclusive AI?

What is AI Ethics

AI Ethics principles give a framework for people working together to have a common language to address ethical aspects of building AI. The top three AI Ethics principles are fairness, trust and transparency. Fairness is about diversity and inclusion in ensuring that the AI treats everyone the same way. Fairness is about AI being free of data bias and algorithmic bias. Data bias[6] is when the data that trains the AI is biased and does not have equal representation for all segments of the population or the data veracity is questionable about certain segments of people. Data bias makes the AI biased against some segment of users. Algorithmic bias is when algorithms make decisions that are biased against certain segments of users. If the AI is self-learning such as a deep learning model that is trained from data, it can become biased by the inherent human bias carried forward in the data. Some reasons algorithms can be biased by biased labeling, bias in human choices made in guiding the algorithm to make decisions or hidden proxies or incorrect bias mitigation methods[7].

[6] Prabhakar Krishnamurthy. Understanding Data Bias, Sept 2019: https://towardsdatascience.com/survey-d4f168791e57

[7] Nicol Turner Lee, Paul Resnick, and GenieBarton. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. 22-05-2019 https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/



Training data are collected and annotated

Human Biases in Data

Reporting bias
Selection bias
Overgeneralization
Out-group homogeneity bias
Stereotypical bias
Historical unfairness
Implicit associations
Implicit stereotypes
Prejudice
Group attribution error
Halo effect

Human Biases in Collection and Annotation

Sampling error
Non-sampling error
Insensitivity to sample size
Correspondence bias
In-group bias
Bias blind spot
Confirmation bias
Subjective validation
Experimenter's bias
Choice-supportive bias
Neglect of probability
Anecdotal fallacy
Illusion of validity

Image: Data Bias ref Lionbridge.ai from Susanna Raj's Inclusive AI Masterclass[8]

Trust is typically synonymous with respecting user's privacy and is tied to data privacy which is governed by GDPR laws and compliance in the EU. There is plenty of research about building human centered interfaces that earns the trust of users[9]. This leads to the concept of Trustworthy AI. Trustworthy AI does not necessarily mean ethical AI.

Transparency is about letting the user know how the underlying AI makes decisions. This leads to the developing field of explainable AI.Transparency is typically talked about as algorithmic transparency and that aspect of AI algorithms is called explainability. Most AI algorithms are built as a black box of thousands of layers of neural networks that deliver a point decision. Researchers are debating about making AI models interpretable with transparency without being a black box which is the practice in most commercial AI available from industry[10]. (add red citation of HDSR article I have added the link at the bottom of references). Some examples of such an AI's point decision are about how an AI recognizes a person from a photo of a face or makes a recommendation for an auto-correct or up-sells some purchase on an ecommerce site.

What Is inclusive AI

All AI is trained using data in the form of numbers, images, videos, sound files or spreadsheets. If this training data is biased and does not include data representing all people, of all cultures, in all conditions, images in all lighting of diverse lived experiences then that AI is not inclusive. Such an AI will be harmful in making decisions when dealing with people and situations that it has not been trained on. An autonomous vehicle that is trained to see bridges in the road will get confused when it comes across a bridge and cause collisions. A non- inclusive AI will discriminate against people who have been left behind in its training data. Inclusive AI treats everyone equally and is inclusive of all people of all races, genders, ethnicities, and every possible perceivable differences and lived truths. For example, an inclusive AI should treat people of all races, people wearing glasses, or people with beards the same way. An inclusive AI should also have clearly defined constructs for abstract words like attentive, movement etc in order to provide a consistent, logical definition for an AI to be inclusive of cultural nuances and situations of lights, sounds, and other environmental conditions.

[8] Susanna Raj, Inclusive AI course at Business School of AI https://businessschoolofai.teachable.com/courses/aiethicscourse-susannaonly/lectures/34275444

[9] 2020. Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery, New York, NY, USA.

[10] Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d

IV. How is AI Ethics and Inclusive AI related to the Autonomous Vehicles

AI Ethics principles as it relates to Autonomous Vehicles:

Transparency:

(i) ADAS features in Level 3 autonomous vehicles make decisions on behalf of the human to keep them safe. For example, “Forward Collision" ADAS feature uses proximity sensors to check if the car is getting too close to the vehicle in front of it. The AI then decides to alert the driver or turn on adaptive cruise control to automatically reduce the speed of the vehicle to avoid collision. There is no transparency about the AI that powers these decisions inside the vehicle.

(ii) In tracking alertness of humans, some car suppliers are offering Infrared sensors on the rearview mirrors[11] to watch the retina movements of people to check for alertness. This is not communicated to people to get their consent or respect their privacy of being monitored inside their own vehicle.

Trust:

People tend to trust Machines and AI more than they trust humans[12]
Driver attention features in Level 2 and Level 3 vehicles have an AI in the vehicle that decides if a human is detected, and whether they are alert. In Level 2 they alert the human and in Level 3 the AI takes control to initiate adaptive cruise control to reduce speed or bring the car into the lane. This becomes critical in the rare situation that the AV loses control and wants to disengage and give control to the human. In that case the AI identifies that a human is present and decides that the human is alert enough to take control. Only then the AI gives control to the human. Today, humans seem to trust the AI to give them control with adaptive cruise control and do not not suspect that the AI might not give them control because of any biased assessment.

The Driver Attention AI model is the focus of this paper because first, the AI has agency on making the decision if and when to give control. Second, the AI is biased and not inclusive.Third, the AI is designed with an ambiguous construct called “alertness” which simply looks for the person’s eyes to be open as a sign of not sleeping and hence alert

Inclusiveness of AI in the Vehicle

Inclusive AI is important in general to treat all people equally wherever AI interacts with humans. The need for Inclusive AI is also unique in some ways different from when AI is not inclusive in non vehicle situations.

In Autonomous Vehicles it is doubly important that AI be inclusive because when an AI is not inclusive, it typically affects the people it is biased against. But with autonomous vehicles, non-inclusive AI increases the liability on the person in the

[11] https://www.bosch-mobility-solutions.com/en/solutions/assistance-systems/driver-drowsiness-detection/

[12] 2020. Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery, New York, NY, USA.

vehicle who is not recognized by the AI but also hurts people and property on the road from accidents when the human in the vehicle is not given control because of a bias from the AI.

In Level 3 Autonomy, the car has to give control to the human driver when it disengages when the AI in the car is unable to drive autonomously. When it does this, the AV will check if the human in the vehicle is awake, alert and then only give control. The vehicle has full agency in making this decision when it is unable to drive safely. First this is flawed in that the human has no agency in this decision.

For example, there was an Uber accident where the car plowed into a fire truck that was reversing into the road in front of the vehicle and people nearby screamed at the vehicle to stop and of course, being autonomous it was not not trained to listen to people. A human driver would have listened to people even if the AV tech got confused by a partial view of a fire truck diagonal view in front of it on the road. In a Level 3 autonomous vehicle, the vehicle passes control to the human.

What does bias with non-Inclusive AI in handing over control to a human mean?

The AI in the car looks at the person in the car using cameras and uses computer vision to judge whether the human is alert. AI is inclusive and ethical if it will give control to all humans equally. It is biased if it has low confidence in judging whether a human is alert based on their looks, race, if they have a beard or their clothing or anything else.

The AI decides if the human is alert based on what training data it has used. The AI is also trained by what is the construct of alertness. Today the construct of alertness is not defined based on cognitive science research and is loosely defined as a person who is awake as seen by eyes being open is alert. Otherwise they are sleeping and therefore not alert. This construct of alertness is flawed and we show by research several situations where the person's eyes may be momentarily closed and they may still be alert. On the other hand, a person whose eyes are open may still not be alert to take up driving of a vehicle impacting the safety of people

IV. Experiment/Research Findings

EXPERIMENT 1: Facial Detection of a diverse set of faces to test for detection faces without bias in the car

Table 3: FairFace Dataset

DATASET: We used FairFace Dataset[1] 86K images in a training dataset balanced for 7 ethnicity, gender and age. FaceFace

Fairface dataset	White		Black		East Asian		South Eastern		Latino		Indian		Middle Eastern		Total	
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
Sample size	7826	8700	6136	6096	6141	6146	5183	5612	6715	6652	5909	6410	2847	6369	40757	45985
Gender %	19.20	18.92	15.06	13.26	15.07	13.37	12.72	12.20	16.48	14.47	14.50	13.94	6.99	13.85	46.99	53.01
Total %	9.02	10.03	7.07	7.03	7.08	7.09	5.98	6.47	7.74	7.67	6.81	7.39	3.28	7.34		

EXPERIMENT DETAILS : FairFace researchers did Face detection rates of commercial APIs on FairFace dataset for Amazon, Microsoft, Face++ and IBM Face detection algorithms. We added the face detection rates on the same dataset using Google Cloud Vision[13] and got detection confidence.

Table 4 : Face Detection Confidence Threshold for Level 3 Control HandOver for Men

Threshold set: Confidence level	**White**	**Black**	**East Asian**	**South Eastern**	**Latino**	**Indian**	**Middle Eastern**
> 90	36.3%	22.29%	28.52%	28.33%	34.23%	26.44%	34.87%
> 80	73.31%	59.47%	69.82%	68.71%	73.2%	65.38%	73.84%
> 70	88.55%	80.99%	88.38%	88.01%	89.42%	85.71%	89.83%
> 60	95.7%	90.62%	96.1%	95.69%	96.06%	95.27%	95.78%

[13] Google Cloud Vision API https://cloud.google.com/vision

If the AV company chooses to set the threshold confidence level at 90%, then the risk of identifying the human is low. This leaves behind people who could take over control in an emergency disengagement and save lives and property. This risk gets exacerbated for people of color as the number of people who get detected goes down.

Today, what confidence threshold the AI operates in is not transparent.

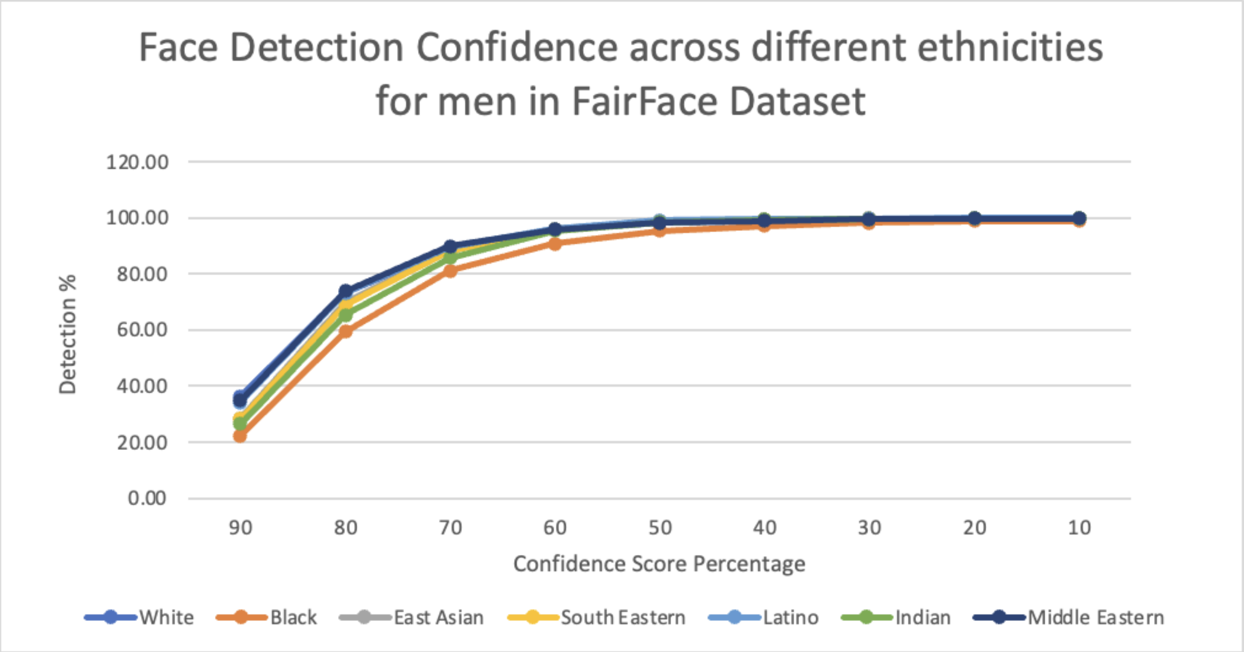
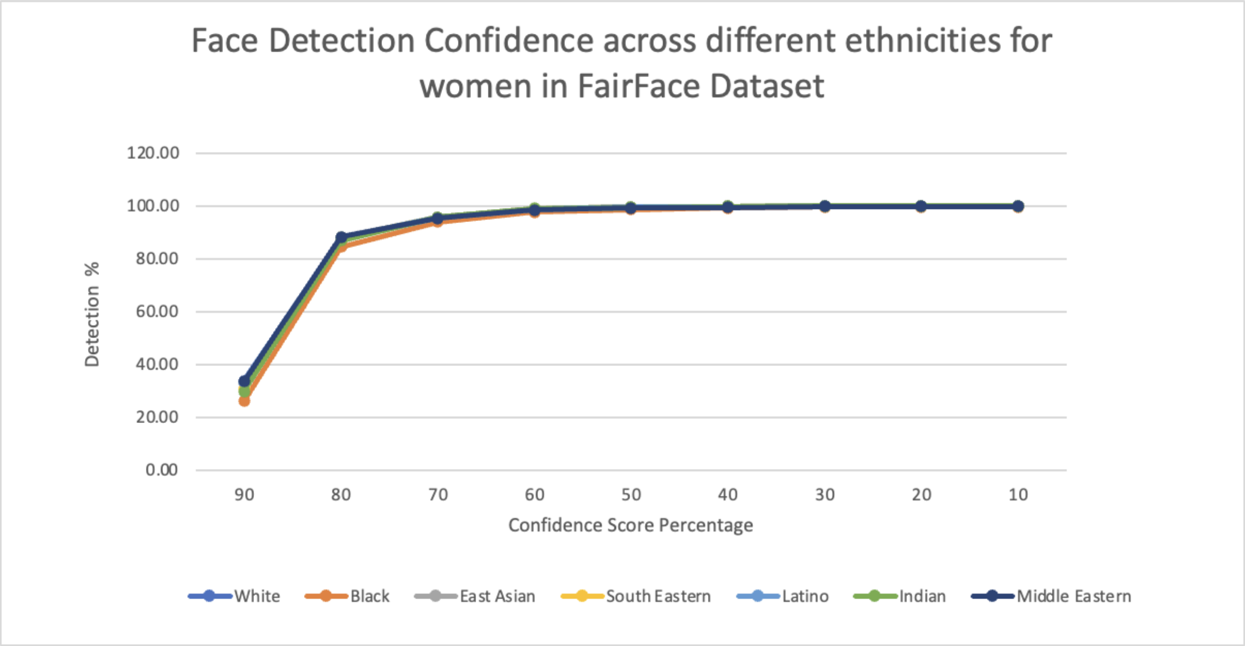


Table 5: Face Detection Confidence Threshold for Level 3 Control HandOver for Women

% of data identified is shown for each ethnicity for each confidence level chosen.

Detection Confidence	White	Black	East Asian	South Eastern	Latino	Indian	Middle Eastern
> 90	33.7%	26.12%	33.07%	30.7%	33.31%	29.73%	33.44%
> 80	87.45%	84.53%	87.72%	87.96%	88.12%	86.77%	88.13%
> 70	95.35%	93.89%	95.52%	95.58%	95.74%	95.75%	95.19%
> 60	98.13%	97.57%	98.7%	98.88%	98.78%	98.93%	98.38%

It was observed that women were detected with higher confidence than men. The chart is for women of various ethnicities.



Findings:

AI detects the human faces in the vehicle and as seen from the table, it has varying degrees of confidence in recognizing a human face of different people. It shows low confidence for people of certain ethnicity, gender and age. People who design the In-vehicle experience decide what level of confidence of the AI is acceptable in allowing the AI to reject some humans. This is not transparent to the outside world even within the company. It is possible that some of these are not human faces and maybe mirage or reflections. But some people may be affected by the bias or non-inclusive nature of the AI. This becomes important when the AV tech fails and wants to give control to the human and uses the AI to decide if there is a human in the car. The people not recognized by the AI in the driver alertness detection system built upon this AI will not be allowed to take care in such situations.

Typically outside of the Autonomous Vehicles, in other applications of AI, bias affects only the people who are marginalized and not included in the AI training data. For example, facial recognition systems are known to be biased against black people and fail to recognize them in Phones or airport checkpoints which use facial recognition.[14] In such cases the bias creates inconveniences and reduces opportunities to the people who are not recognized by the AI by failing to

[14] Buolamwini, J., Gebru, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

identify them or misidentifying them. Cathy O’Neil in her book “Weapons of Math Destruction"[15] shows how women and minorities are not given equal opportunities in hiring and loan approvals because of bias in AI. But in an autonomous vehicle this is different. When the AI in the vehicle is biased it will not recognize the person. This is different from identifying a person to give them some place because in this case the AI driver management system needs to identify a person when it needs to give control to the person to drive when the AI fails and needs to disengage. Failure to recognize the person will affect their safety but also will also hurt the safety of the public on the road because the human who is not recognized is not given control of the vehicle to drive it to safety, thereby letting the AI fail and create a collision. This can hurt any number of people and property on the road. This is a much bigger issue than non recognizing one person inside the vehicle and is the core of our research to highlight this issue with a recommendation to mitigate the risk while working with a biased AI.

As seen from the image above, if an AV tech company chooses a confidence of 60% to allow face detection to identify an alert driver to give control and another AV tech company chooses a confidence of 90% to allow face detection to identify an alert driver to give control, the difference is not just annoyance of people of certain ethnicity impacted by the bias. It is a missed opportunity to allow control to more people who might be able to take control of the vehicle in an emergency and save lives on the public roads.

Without consideration of AI Ethics principles of fairness, a set of people are distrusted by the AI and without the AI Ethics principle of transparency, AV vehicles companies make this level of confidence as an arbitrary decision point and impact lives of the public on the road.

This raises the moral aspect of saving lives not just by fairness and inclusive AI but by transparency of decisions about the level of confidence of the AI which can shift the blame for a collision to people in the vehicle for not taking control or the AV tech for not giving control to certain people who could save lives and property. This is going to have an important impact on insurance risks and claims in the future too.

Recommendation from experiment 1:

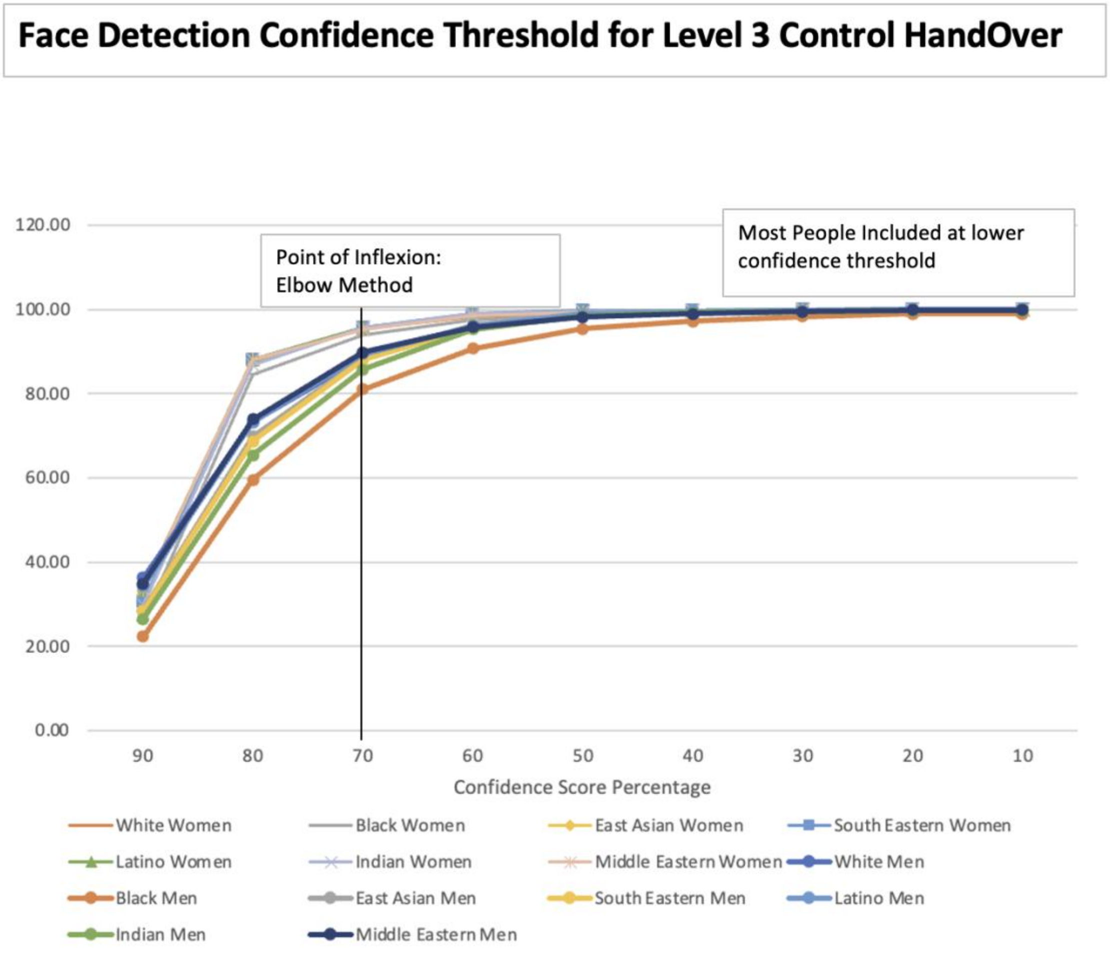
Our recommendation is to follow the elbow method of clustering heuristics used in AI clustering algorithms and plot an elbow diagram of the entire data and capture the elbow point of confidence threshold at the optimal point of diminishing returns to determine the confidence threshold to set. This will ensure the maximum number of people will be detected by the AI in the vehicle to handover control.

Note, this is good for the vehicle handover use case and will work contrary to any other AI in the vehicle that is trying to predict humans to enable better experience or infotainment in the car

The selection of inflection points raises a serious concern about the current model under use in this research. Our recommendation is to use a better balanced

[15] “Weapons of Math Destruction” book by Cathy O’Neil. https://en.wikipedia.org/wiki/Weapons_of_Math_Destruction

dataset similar to FairFace and train vision models to be bias free. This exercise improves the models for developers to use in their applications and hence mitigate the bias.



Face Detection Confidence Threshold for Level 3 Control HandOver

EXPERIMENT 2a: Alert construct definition is flawed with limited definition of alertness. We tested for different actions to test if the face was identified for sneezing/yawning/crying faces.

AVFACE DATASET: We created a new set of 30 images of a diverse set of faces of people of multiple ethnicities, gender and age with 10 images of crying faces, 10 images of yawning images and 10 images of sneezing images. We call this dataset the AVFACE Dataset.

AVFACE DATASET

Experiment Details: Alertness in Autonomous Vehicles is detected as eyes being open as a simplistic definition. So we wanted to prove that sneezing, crying and yawning can create situations where the person is alert but eyes are not open.

First we did simple face detection on AVFACE dataset to see if yawning, crying and sneezing makes the face not detectable. We used Microsoft Face API, Amazon Rekognition and IBM's Facial Recognition all well known APIs but currently on a moratorium as they are all banned for Police.[16] But Microsoft and Amazon continue to offer it for testing without any conditions on their site. [17]

Table 6: Yawning Faces Detection:

Commercial Face Detection API used	**Face Detection Confidence on AVFace Data**				
	Not Detected	**01- 30**	**31- 60**	**61-90**	**91-100**
Google Cloud Vision API	0	0	20.00%	50.00%	30.00%
Microsoft Face API	7.69%	0	0	0	92.31%

Table 7: Sneezing Faces Detection:

Commercial Face Detection API used	**Face Detection Confidence on AVFace Data**				
	Not Detected	**01- 30**	**31- 60**	**61-90**	**91-100**
Google Cloud Vision API	0	10.00%	20.00%	40.00%	30.00%

[16] Moratorium on Face detection API by US Government
https://www.forbes.com/sites/larrymagid/2020/06/12/ibm-microsoft-and-amazon-not-letting-police-use-their-facial-recognition-technology/?sh=292e0f918871

[17] Microsoft Face API https://docs.microsoft.com/en-us/azure/cognitive-services/face/concepts/face-detection

Microsoft Face API	0	0	0	10	90

Table 8: Crying Faces Detection:

Commercial Face Detection API used	**Face Detection Confidence on AVFace Data**				
	Not Detected	**01- 30**	**31- 60**	**61-90**	**91-100**
Google Cloud Vision API	15.38%	7.69%	15.38%	38.46%	15.38%
Microsoft Face API	7.69%	0	0	0	92.31%

Findings:

Google Cloud Vision API was able to detect all the faces that were yawning or sneezing faces. It failed in recognizing exceptions in recognizing crying faces. Microsoft Face API[18] on the other hand had a higher confidence in recognizing most faces but had exceptions for unkempt beard faces.

Exceptions:

1. Microsoft Face API did not recognize the crying face of a bearded person with a bias towards unkempt hair while it worked for a trimmed beard face (below). It identified it as "dog lying on its back."

2. Google Cloud Vision API was biased without cultural sensitivity to women hiding their face when crying and it failed to recognize it as a human face.

Recommendation: Humans when left alone in the vehicle when the car drives autonomously are not going to sit with a straight face. They will be laughing, sneezing, crying, yawning, talking, gesturing, and doing more actions. So the AI models should be trained with faces showing multiple actions for it to have a good performance in identifying the human in the vehicle in real time.

[18]

EXPERIMENT 2b : Alert construct definition is flawed by defining alertness as eyes open.

Results 2b: Prevalent alertness construct is tracking eyes open as alertness which is flawed. We ran **Amazon Rekognition API** on AVFACE dataset to check for eyes open and it showed the alert construct is flawed.

Table 9: Eye Open Detection on AVFACE dataset:

Commercial Face Attribute Detection API	**EyesOpen**		**EyesClosed**	
	Confidence > 90	**Confidence < 90**	**Confidence > 90**	**Confidence < 90**
Amazon Rekognition API	15.62	25	21.87	37.5

Experiment 2c: Eyes Open detection to check for Alertness is biased for different ethnicity and gender. We ran Amazon Rekognition API on FairFace Dataset (attribution) 86K images in training dataset. It is biased. Google Cloud Vision does not have an eyes open option.

Table 10: Eye Open Detection on 30% of FairFace dataset by ethnicity

	White		Black		East Asian		South Eastern		Latino		Indian		Middle Eastern	
	F	M	F	M	F	M	F	M	F	M	F	M	F	M
Amazon Rekognition API	0.78	0.79	0.8	0.78	0.72	0.68	0.76	0.75	0.81	0.81	0.79	0.84	0.74	0.79

Findings:

1. As seen from table 9, only 15% of the faces in the AVFACE dataset of people’s faces in action - crying, yawning or sneezing - are detected as eyes opened. That means 80% of the people are considered to be not alert and hence the AV will not give them control in an emergency situation. This is a huge risk because people will be showing some action

on their face and not sit idle staring at the road when they are not driving in an autonomous vehicle. This creates a huge risk because of the limited construct of alertness defined in a primitive fashion as eyes opened.

2. As seen from table 10, we can see that the model is biased against asian faces in recognizing eyes as open. This impacts the safety handover to alert people who could save the vehicle and also protect people on the road from a collision.
3. It is not known what API the companies are using for eyes open detection to claim alertness. Transparency is needed in order to trust the algorithms.

Call for Collaboration

We call for collaboration from researchers interested in making the autonomous vehicle inclusive and safe by reducing bias in the AI in the vehicle.

Further research is needed to improve the construct of alertness in the vehicle to include all kinds of actions displayed by humans in real-time. We need collaboration from researchers from around the globe to ensure that we have data showing a variety of ethnicities, cultural nuances of actions of crying, yawning etc. Given that the AI fails in recognizing beard faces, we should expand that to include faces with glasses, faces wearing masks and more.

We could expand the inclusive AI in the vehicle to go past driver alertness to check for other AI in the vehicle that interacts with humans using voice or AR and or created infotainment experiences.

The promise of autonomous vehicles is to save lives. We hope that we can collaborate with other researchers to ensure that we make the AV inclusive so it can focus on improving its autonomous capabilities to save lives.

Conclusion

We set out to find if the AI inside the autonomous vehicles that gives control over to humans is inclusive and treats all people the same. We found several AI Ethical issues 1. The AI is not transparent on what sensors are tracking humans and what the cameras are watching or making deductions about humans without any consideration to their privacy. 2. We found the face detection used in the cars is similar to facial recognition algorithms used elsewhere they are banned in cities and for use by police and there is no restriction for us inside the vehicle and our research showed that the datasets are biased and the algorithms are trained by biased datasets and are unfair to people of certain ethnicity, gender and age. One shocking observation was that such a bias does not only affect the people undetected as humans in the vehicle but by refusing to give them control when the AV disengages, it creates a huge risk of collision and will hurt people and property on the road. We recommend the elbow method in finding the optimal confidence threshold to ensure the maximum number of humans are identified irrespective of the performance of the AI model in detecting humans.
3. AV companies use a construct for defining alertness as just checking if a person’s eyes

are open. We created a new “AVFace dataset” with faces of people yawning, crying or sneezing and showed that this construct is flawed in detecting alert humans ready to take control of the vehicle.

Our recommendation is to make the face detection transparent and trustworthy by reducing bias and expanding the alertness construct beyond eyes being open but more importantly, go beyond the algorithm in creating human centered design inside the vehicle. Today the machine has agency to decide when and who to give control to humans in the vehicle. This AI not being inclusive ethical AI affects the humans in the vehicles and also harms people and property on the road. So our top recommendation is to change the design to give agency to the human to take control and for the AI to improve inclusiveness and augment human decision when a human is present in the vehicle. Our aspiration is for an ethical inclusive AI to augment humans inside the vehicle to keep them safe in the cars and on the roads so autonomous vehicles can attain their promise to make our roads safe.

References

1. Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558).
2. Wilson, Benjamin, Judy Hoffman, and Jamie Morgenstern. "Predictive inequity in object detection." arXiv preprint arXiv:1902.11097 (2019).
3. Nyholm, Sven, and Jilles Smids. "The ethics of accident-algorithms for self-driving cars: An applied trolley problem?." *Ethical theory and moral practice* 19.5 (2016): 1275-1289.
4. SAE International Releases Updated Visual Chart for Its “Levels of Driving Automation” Standard for Self-Driving Vehicles, 2018-12-11 WARRENDALE, PA. https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles
5. “2030 The Driverless World: Business Transformation from Autonomous Vehicles”, Sudha Jamthe, Sep 2017 (https://www.amazon.com/gp/product/1973753677/)”
6. Prabhakar Krishnamurthy. Understanding Data Bias, Sept 2019: https://towardsdatascience.com/survey-d4f168791e57
7. Nicol Turner Lee, Paul Resnick, and GenieBarton. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. 22-05-2019
8. Susanna Raj’s Inclusive AI Masterclasshttps://businessschoolofai.teachable.com/courses/aiethicscourse-susannaonly/lectures/34275444
9. 2020. Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery, New York, NY, USA.
10. Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d
11. 2020. Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery, New York, NY, USA.
12. https://www.bosch-mobility-solutions.com/en/solutions/assistance-systems/driver-drowsiness-detection/
13. Google Cloud Vision API https://cloud.google.com/vision
14. Buolamwini, J., Gebru, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency
15. “Weapons of Math Destruction” book by Cathy O’Neil https://en.wikipedia.org/wiki/Weapons_of_Math_Destruction
16. https://www.forbes.com/sites/larrymagid/2020/06/12/ibm-microsoft-and-amazon-not-letting-police-use-their-facial-recognition-technology/?sh=292e0f918871
17. https://docs.microsoft.com/en-us/azure/cognitive-services/face/concepts/face-detection