Algorithm Audit

Focus group

Predicting irresponsible driving behavior

Algoprudence

ALGO:AA:2025:01:FG

13 januari 2025

Executive summary

For the case study 'Predicting irresponsible driving behavior', Algorithm Audit convened a focus group. The group consists of four individuals who regularly use shared vehicles, such as scooters, bicycles, e-scooters and cars, through platforms like GoSharing, Felyx, Check, Tier, GreenWheels, Sixt and MyWheels. Participants were approached through the professional network of Algorithm Audit. The focus group's purpose is to gather users' insights on data collection, algorithmic predictions and communication methods used by shared mobility platforms.

The insights gathered from the focus group serve as input for the normative advice commission, convened by Algorithm Audit, which will offer guidance on the socio-technical questions raised in the problem statement document. The case study described in this document focusses on the calibration of a machine learning-based risk profiling algorithm designed to identify inappropriate driving behavior within the database of a business-to-consumer car-sharing platform. The full description of the case can be found in the problem statement document 'Predicting irresponsible driving behavior' (ALGO:AA:2025:01:P).

About Algorithm Audit

Algorithm Audit is a European knowledge platform for AI bias testing and normative AI standards. The goals of the NGO are four-fold:

Knowledge platform

Bringing together experts and knowledge to foster the collective learning process on the responsible use of algorithms, see our white papers and public standards.

Normative advice commissions

Advising on ethical issues that arise in concrete algorithmic practice through deliberative and diverse normative advice commissions, resulting in algoprudence

Technical tools

Implementing and testing technical tools to detect and mitigate bias, e.g., sociotechnical evaluation of generative AI, unsupervised bias detection and synthetic data generation.

Project work

Supporting public and private sector organisations with specific questions regarding responsible use of AI, from a not-for-profit perspective.

1. Practical details

The focus group was convened on 13-07-2024 to learn more about users' views on data collection, algorithmic predictions and communication by shared mobility platforms. Participants were approached through the professional network of Algorithm Audit representatives and consisted of in total four people: 2 men, 2 women, aged between 20-30 years old. The in-person session was held in Almere and moderated by a representative of Algorithm Audit. A series of questions were posed to gather insights on users' experiences with shared mobility platforms, as well as their thoughts on the role of data analytics in this context, which led to group discussions. The questions focused on data collection, monitoring driving behavior, detection of irresponsible behavior, the process for blocking accounts (including human interaction and appeal), and the trade-off between false negatives and false positives.

After the session, the group interview was transcribed and summarized into this report. The answers by focus group participants are transcribed as directly as possible, with efforts made to retain the original tone and to translate them into English as accurately as possible.

When the focus group was convened, the number of linked bank accounts to a platform user, traffic fines and leaving the car dirty were still variables used in the risk prediction model, and are therefore addressed in some questions. In the meantime, the platform decided to exclude these variables from the risk prediction algorithm. More details can be found in the problem statement document 'Predicting irresponsible driving behavior' (ALGO:AA:2025:01:P).

NOTE: The insights from this focus group do not necessarily reflect Algorithm Audit's perspective on responsible data processing and the responsible use of algorithms.

2. Summary of answers

In total 20 questions were raised to the focus group participants. Questions were asked across four themes: introduction, data collection, algorithmic risk prediction and actions following predicted irresponsible behavior. The following is a summary of the responses to questions about these topics. All questions and answers can be found in the section Full meeting minutes.

2.1 Data collection

The customers of shared mobility platforms who participated in the focus group perceive data collection of rental behavior as important. They particularly emphasize the value of standardized driving data for identifying irresponsible driving behavior. The participants perceive data collection to be ubiquitous and unavoidable. However, they are not always aware of the fact that their data is being collected and for what purpose. The participants consent for data processing when starting to use the services of the platform, but the exact data processing practices that are described in the privacy statement were rarely reviewed by them. Participants of the focus group believe that key data points for assessing irresponsible rental behavior include fuel filling behavior, vehicle damage, gear shifting patterns, issues with lights or batteries, years of driving experience, and whether the account of someone else has been used. Participants of the focus group consider payment behavior also to be relevant to analyze irresponsible rental behavior, especially for users who appear to be

concealing something or have a history of late payments. They find that the following data points could be considered to identify irresponsible payment behavior: foreign accounts while being registered at a Dutch address, whether a payment card actually belongs to a user, unpaid fines, the use of credit cards, refused direct debits, and income.

All questions raised and provided answers can be found in section 2. Data collection.

NOTE: These suggestions do not necessarily reflect Algorithm Audit's perspective on responsible data processing and the responsible use of algorithms.

2.2 Risk prediction algorithm

The customers of shared mobility platforms who participated in the focus group consider it important that irresponsible users are warned by the platform before being blocked. They find that data analytics can be used to identify characteristics of irresponsible behavior based on historical data of blocked customers, such as minor violations[1], e.g., returning the car 5 minutes later than the pre-determined time, unreported damages or a dirty car. Users can actively help with identifying irresponsible behavior by reporting violations of a previous user in the app of the platform. Participants of the focus group believe that monitoring driving behavior of all customers increases individual trust in the platform's services, as the downstream effects of irresponsible driving behavior of others on them is reduced.

The participant prioritizes that as many irresponsible drivers as possible are identified, i.e., prefering a small number of false negative predictions even if this results in more false positive predictions. This is because safety is jeopardized when irresponsible users go undetected. However, the participants find that tolerating too many false positives brings the risk of penalizing users who are behaving responsibly, with possible multiple unnecessary consequences attached to it (such as a blocked account, decrease of platform loyalty by users and surge of appeal procedures). As a result, shared mobility platform risk losing responsibly behaving users, especially considering how easy it is to switch to another platform and continue using similar services. This is an unpleasant outcome for the platform, because the users are being chased away who otherwise would have kept using the platform's services.

The participants shared multiple ideas regarding the detection of irresponsible behavior. For instance, obtaining camera footage as a method of verifying irresponsible behavior of users. They found it acceptable that cameras on the exterior of a vehicle could be installed to monitor and verify if the user adheres to the rules properly. Besides, customers should always have the opportunity to appeal the decision of their account being blocked, so that unjust decisions can be rectified. Furthermore, it is imaginable that irresponsible users continue their risky behavior by using the services of other shared mobility platforms. To prevent this, participants suggested a solution to create a shared database that all vehicle platforms can access in order to have an overview of irresponsible users and prevent them from carelessly using the services of the platform.

All questions raised and provided answers can be found in section 3. Algorithmic risk prediction.

[1] More information about the precise process followed for data analysis can be found in the problem statement 'Predicting irresponsible driving behavior' (ALGO:AA:2025:01:P).

NOTE: These suggestions do not necessarily reflect Algorithm Audit's perspective on responsible data processing and the responsible use of algorithms.

2.3 Actions following predicted irresponsible behavior

The participants believe that there must be an opportunity to appeal a decision before an account is blocked. A permanent decision should only be made after obtaining all the relevant facts. In this respect, the focus group participants find it important that contact between the user and the platform is possible in case of a suspicion, preferably via email, telephone and/or video call. A platform representative should be someone who can properly assess whether a prediction by the algorithm has been made correctly and who handles the compliant in an unbiased manner. The participants find that the manual check should not be replaced by a fully automated decision-making process of blocking someone's account.

Furthermore, focus group participants find unnotified blockage of an account after detecting irresponsible behavior, manual or algorithmically, undesirable. The preferred method would be a layered approach to address irresponsible behavior, in which the user gets a warning before being permanently blocked, for instance by imposing first a temporary block of for instance 1-7 days. However, for serious matters, such as excessive speeding, the participants find a longer or permanent blockage justified in the light of safety. Besides, rather than solely punitive actions based on data analytics, focus group participant find that platform customers should also be rewarded for good behavior. Examples can include discount codes, a loyalty system, positive reviews from fellow customers, or a credit score system.

Lastly, focus group participants customers consider it to be important that they are informed in what manner a decision was made and what role the algorithm played in the decision-making process, which should be explainable to the customers afterwards. This includes transparency about the driving and payment behavior data used for algorithmic predictions and personal data of customers used for manual inspection before blockage.

All questions raised and provided answers can be found in section 4. Actions following an algorithmic prediction.

NOTE: These suggestions do not necessarily reflect Algorithm Audit's perspective on responsible data processing and the responsible use of algorithms.

3. Full meeting minutes

In this section, the meeting minutes of the focus group customers are presented. Across four sections, all 20 raised questions and provided answers are stated. First, the introductory questions are discussed, then questions relating to data collection, algorithm and datapoints, and methods of evaluating the algorithm.

3.1 Introductory questions

Question 1: Which shared mobility platforms are you familiar with?

Check, GoSharing, Felyx, Tier, MyWheels, Sixt and Greenwheels.

Question 2: Do you ever use shared mobility?

Yes, some participants frequently use a shared e-scooter (Dutch: elektrische scooter) or scooter (Dutch: step). Also, some participants use a car through a car sharing platform, usually once a month or once every two months.

Question 3: Why do you use a shared mobility and/or car sharing platform?

It is especially practical when there is construction work or maintenance for public transport services and you suddenly cannot take the bus, train or other services. It is also useful if you are abroad and want to rent a car. One of the participants does not have a car themself and finds it useful to drive a shared car for long distances, especially if public transportation takes too long or there are (sudden) delays.

3.2 Data collection

Shared mobility platforms collect data on your driving behavior to analyze which aspects of that behavior align with the behavior of users who have been blocked for irresponsible driving.

Each bullet point represents a comment made by a focus group participant.

Question 4: For this purpose, individual driving behavior is collected as data, what do you think about that? Such as location, speeding events, sudden driving movements (sharp turns, hard braking, switching lanes without signaling).

> The fact that driving behavior is collected is good, such that a platform can inspect whether people are responsible drivers. It is an appropriate way to keep track of which users, for example who do not have a car themselves, break the rules, such as driving at high speed or use a car at a particularly late hour.
> Collecting data is part of the service, but there is certainly some reluctance to inform users properly that personal data are being collected.
> It makes sense that they collect personal data, but it does imply that there are things in the car or that you are being looked at which you are not aware of. To protect your own safety and the safety of others, it is important that data is collected in this manner though.
> It should be taken into account that your driving behavior also depends on driving behavior of other users. For example, in the case that you have to stop because someone brakes in front of you, if that behavior is negatively linked to your account, it is especially unfair because you are responsible for someone else's mistake.

Question 5: And payment behavior is also collected, what do you think about that? Such as the amount of registered bank accounts.

> It is not wrong that payment data is collected. You can assume that a person has something to hide or is suspicious when more bank accounts are specified.
> Participants use PayPal or direct debit because it is easier. Suppose someone is not aware of a payment deadline, there are no consequences because that gets deducted from your account automatically anyway.

> Nowadays you cannot avoid that such behavior is monitored, it is better that it is something small like a bank account than other in-depth personal information.
> Future users should not actually notice any effects of the behavior that you display as a bad user (e.g. just not having enough money in your account). It would be unfair if the entire algorithm that uses historical data made generalizations of everyone in the future. In principle, discrimination can then take place on the basis of the past, especially if you as a user simply have all the means to pay.
> Processing payment behavior as data and keeping track of that sounds logical, also in connection with the criminal flow of money. If you do not do anything strange yourself, you will not have to worry about any bad consequences.

Question 6: Are there any specific data mentioned that you agree or disagree with being collected and why?

> With regard to safety, it is very important that the relevant data are collected. For example, irresponsible behavior, such as returning a vehicle too late or exceeding the speed limit, says something about certain users.
> If you create damage or are responsible for an accident, this can have severe consequences even when it is not your fault. This should be represented well in the data collected.
> Driving behavior, particularly irresponsible behavior, should be linked to age for some people, e.g. youngster who often accelerate fast in urban environments.
> With some rides you can unexpectedly get a fine. For example, when a scooter cover is broken by a previous user. Even if your trip is very short, you can still be charged for the condition of the scooter while the previous user is responsible for this. Data points do not always capture the right information. That can cause confusion about inappropriate behavior and who was responsible for it.
> With respect to payments, sometimes the app of a e-scooter platform does not work properly. It happened once that a payment had already been made but the e-scooter would not start and remained in the same place all day. The user could not indicate that the ride had never started and was subsequently sent a bill of 64 EUR.

3.3 Algorithmic risk prediction

An algorithm is trained on a subset of the previously discussed data to inform users about their driving behavior and how it can be improved.[2] The user receives a notification if they show signals of irresponsible driving behavior. When the algorithm predicts that a certain user is showing irresponsible driving behavior, a user can be blocked by a human analyst of a shared mobility platform, based on the risk prediction made by the algorithm. Depending on how high the risk score is, either no action is taken, or you receive a warning or you are blocked.

Each bullet point represents a comment made by a focus group participant.

[2] The participants were shown an earlier version of all datapoints as listed in Appendix B of the problem statement 'Predicting irresponsible driving behavior' (ALGO:AA:2025:01:P).

Question 7: When an algorithm is used by a car sharing platform to inform users about their driving behavior, in the form of a notification as mentioned, would you use the services or not? Why?

> Safety first. So, if the algorithm is used in that context, I would have nothing against its application.
> Using the algorithm is smart to monitor driving behavior. Suppose you are not wearing your seat belt, then that should be considered whether someone can keep using their account. But when the behavior is linked to someone's migration or personal background, then that is undesirable. It depends on what is done with your data and for what purpose.
> In first instance, using it would be fine, but it also depends on your own driving behavior. If I were to receive such a message often, I would stop using the platform at some point.
> It is very understandable that data analytics is used, because digitalization is important and cannot be ignored, as it is a part of daily life. But before consequences are attached to someone's behavior, it must be explained and clarified why a certain decision was made.
> It is essential that there is still a form of contact between the user and the platform. Suppose something happens, then the platform must question the involved parties about what exactly the situation entailed. If you receive a notification, the user should have the option to call, whether it is at an intermediate step or final step in the feedback or blocking process. Your side of the story is important, especially if you have a good explanation for how something happened. In addition, an employee of a shared mobility platform should be able to properly assess whether someone is being truthful or not and whether the prediction of an algorithm is correct.

Question 8: When an algorithm is used to identify and block irresponsible users, what pros and cons can you think of when using the platform's services?

Advantages: safety always comes first, which means you can use the services with more peace of mind. Trust in the platform increases when you know that users have handled the vehicle properly before you use it yourself. When customers have been using the services for a long time, you can determine which users behave normal and the quality of the shared items will improve. For example, if it is recorded when someone does not refuel and leaves the car in that state, the next user can report that the previous person behaved incorrectly and can therefore also contribute to data collection. If historical data is used from a long period of time, the quality of the platform's rental items can be improved and maintained cost effectively. It is important how the data is used. In addition, it is also assumed that cost efficiency through algorithm use is processed in the price, which can possibly make the services more affordable.

Disadvantages: if you are labelled as irresponsible, but cannot express your point of view. It is not easy to see exactly which data of an individual are used in the algorithm or whether you as a user can easily see that somewhere in an overview, which is inconvenient. A risk score should not result in blocking or a warning all at once, but rather keep track of how irresponsible behavior develops. Then you can take notice of specifically as a consumer: I am doing something wrong, I need to do it differently. Sudden blocking does not help, you need to be able to see the reason why your risk score is going up, for example. Or that a 'three strikes' principle is applied before you are blocked at the platform.

Question 9: Suspicious driving behavior includes, for example: leaving the car behind messy or dirty, not paying on time, and violating traffic rules more often. Or if you place a scooter or bicycle incorrectly on the road, which obstructs other traffic. Which factors, i.e. different driving behavior, should also be taken into account here?

> Fuel filling behavior: with scooters you automatically see to what extent it is charged in advance. With cars this is not really clear. Someone may not have filled the fuel tank in advance, which is annoying for the next user if they did not empty it, but in fact the person before them.
> Leaving the car dirty: this should include smell, cars often do not smell nice. Especially if someone has smoked or something like that. As a platform you could maybe find out who it was. But often the next users suffer the consequences in the moment, not the people who have to clean it themselves (e.g. the previous user). This also involves reluctance to report such problems, because there is a fear that the platform will think that the current user did it precisely because they reported it.
> Certain types of damage: if things have been lost from the car or are broken or not in position as they should be.
> Shifting gears: if you do not shift gears properly, this contributes to a broken car. As a new user you do not notice this yourself when you just get in.
> Lights or batteries that do not start: this is annoying when you are far from home and the previous person has caused the defect. But often they cannot trace the specific person and you do not get your money back for a situation like this.
> Individual violations: sometimes there are 5 people using one account and that can be very dangerous. Especially if you want to find out who is at fault.
> Additional note: it would be useful if the car meets the conditions of a certain checklist before the user enters the car.

Question 10: Are there certain factors that should not be taken into account in suspicious driving behavior?

> Age: this should not affect driving behavior by itself, but it can make a difference in exceptional cases, for instance, how long someone has had their license. So, people who have had their license for longer tend to drive a little more recklessly.
> Driving experience should be taken into account instead of age: how often someone has rented a car shows a certain experience. It makes a big difference whether something goes wrong on your 1300th ride compared to your second ride.

Question 11: Suspicious payment behavior includes, for example: not paying bills on time or if you have registered 10+ bank accounts to pay with. What other factors should be taken into account here?

> Foreign accounts, in combination with a Dutch address
> Verification: especially if it is not you on your debit card
> Outstanding fines
> Automatic payments that do not seem to work
> Credit cards instead of bank accounts
> Checking whether someone has an income at all if they systematically do not pay
> Extra note: users have to pay in advance, if the debit is not immediately charged

Question 12: Are there certain factors that should not be taken into account in suspicious payment behavior?

> Migration background: low income sometimes correlates with migration background. Relating to payment behavior this should not result in discrimination. Regardless of background, you should always be able to use the services of a sharing platform.
> Age: payment fraud can be committed by all ages, but this should not be linked to a specific age in advance. Perhaps age in combination with years of driving experience can be taken into account, but not age itself. Because someone who is 18 years old who rents a shared vehicle and then does not pay is relatively worse than someone of 30 years old who rents and does not pay once.

3.4 Actions following an algorithmic prediction

Question 13: When the algorithm has predicted that the user has a high risk of irresponsible behavior and exceeds the threshold for being blocked, an employee can decide to block the account. What do you think about this process?

> It should be allowed for safety reasons if a user shows irresponsible behavior.
> This should be done with the principle of hearing the relevant parties, so the user should be able to explain themself. Blocking may only be done with an explanation.
> A temporary block of 24 hours, for example, would be the preferred action to undertake. That is different from completely blocking an account. Based on the decision that would follow, it can then be determined whether you will be blocked for longer or that 24 hours seems to be enough.
> It is important that employees must be able to make a correct assessment of whether someone is actually telling the truth. It is justifiable, however, that in certain situations a user will be blocked immediately, such as driving 200+ km per hour on the road.
> When an employee intervenes, it must be ensured that discrimination and prejudice is not at all an issue in the decision to block. This may include factors such as the user's voice on the phone, name or photos that belong to an identity document.

Question 14: If your account is blocked, what data do you think should be available about the decision and the risk prediction algorithm used? For example, how the algorithm generates a risk score, on what criteria the algorithm is trained etc.

> Driving behavior: speeding events or driving unsafely
> How long you have had your driver's license
> The user's entire driving history in the app
> Previous notifications, if you have received a warning or multiple warnings
> Whether you leave the vehicle behind neatly
> Combination of factors to establish a risk score, e.g. can you drive well and do you pay on time
> Payment behavior: whether someone uses a foreign credit card or does not pay on time

Question 15: If your account is blocked, how would you contact the platform to appeal the decision? For example, file a complaint by letter, reach out to customers support through the website or in the app.

> You should have the choice to speak to someone on the phone.
> The platform must call or send an email with the option to schedule an appointment about the decision.
> Calling on the basis of your own initiative would also be desirable in order to share your side of the story about the situation as quickly as possible and to discuss it.
> Also an email should be sent by the platform, possibly after the conversation, to have a written copy of the discussed content. For instance, a notification after the phone call via email that you have everything in writing which was discussed, in case there turns out to be a disagreement. Then you can also create a file for yourself for possible evidence.
> Online options must also apply, but no chatbot.
> The burden of proof that there is irresponsible behavior must lie at the side of the company/platform. Customers should be able to appeal every decision. This could also be done with another institution so that they can at least check whether the procedure at the platform went well or not.
> In such cases, you must always be able to view your own data and how it has been used.
> Additional note: there is a chance that users will simply switch to another platform if their account is blocked. Because multiple platforms offer similar services, you can simply choose not to challenge the decision and just download a new app. It is imaginable that a shared database could help for these types of companies, which tracks driving behavior etc. S,o that people cannot misuse other platforms. On the other hand, your privacy will be limited in the sense that you will be monitored without you possibly wanting this to happen.

Question 16: For the case study under review, various features are considered by the algorithm to detect irresponsible behavior. Do you also see opportunities to use an algorithm to reward users? If so, in what way?

> Come up with a loyalty system so that users are blocked less. Suppose you are blocked, with rewards you will have a reason to want to do something about it to earn your place back.
> Free driving minutes.
> Discount code for 5 good rides, for example if you have parked well, left the car neatly, did not exceed any speed limits and paid on time, among other things.
> Make the services less expensive because it is often a threshold to use it at all, because it adds up if you pay by the hour, for associated kilometers and perhaps also having to pay for the refueling.
> If you have had 10 rides that afterwards your hourly rate becomes lower. If you use it often that it makes a difference to your bill.
> A points system if you do well, but those points often expire quickly, so they should be valid for longer. If you use public transportation easily or do not need a car very often, it is a shame if the points expire quickly. Such platforms are also not for daily use.
> If someone has left the vehicle clean, you can indicate this in the app and then this person can receive a discount based on the opinion of the person after him. So, for example with 100 points you can use 25% discount on the next ride. And then you also immediately have data about the previous person because users themselves indicate the condition of the vehicle in the app.

> Apply a credit score system that decreases in score with irresponsible behavior or increases with responsible behavior. You can also keep track of data in a more targeted way. You can apply a false name in this system of rewards for the previous user so that their identity remains anonymous, but the system can still find out who it is so that they can receive the rewards on their account.

Question 17: How important is it to manually check, by means of an employee, what users do not adhere to the terms and conditions of the platform? Can this process be fully automated? What do you perceive as good and bad about the element of human intervention that is currently applied?

Explanation: Currently, an employee has the authority to decide whether an account should be blocked after the algorithm predicts a link to irresponsible driving behavior. Imagine if this entire process, detection and decision, was managed solely by the algorithm. How do you view that scenario? What are the advantages or disadvantages of keeping the human element in the process as it is now?

> If your situation is examined, there must be a margin of correction, this is currently guaranteed by the human experts, but cannot be captured in computer code.
> An algorithm often has a good sense of estimating the right outcome so that no wrong decisions are made. So, more automation is okay for me.
> The algorithm can be used to make the entire decision, which is faster and more efficient. And only if someone disagrees with their decision and objects can an employee see what has occurred.
> The algorithm gives advice to the employee through the generation of a risk score. The employee should then call the user to explain what is going on and make a decision from there.
> It should not be fully automated, only if the algorithm is a 100% accurate. In the first five years it should not be without employee intervention. The self-learning ability of an algorithm would be desirable, but should only be deployed after a few years so that the algorithm functions as correct as possible based on the proper data and real life situations.
> Possibility to appeal a decision based on the algorithm regardless of the situation: the decision should be made available by telephone and in writing saying that a formal decision will be made according to certain reasons established by the platform which should be explained. The extent to which you want to continue the proceedings of the appeal depends on you and whether you want to invest time and energy in it. Here it applies that the platform employee on the other line must write down everything. And that this can all be done digitally, that you do not have to be at a certain location. Preferably via Zoom or something similar.
> Online conversation would be preferred, face-to-face is a bit more comfortable. As a user, you take it very seriously if you receive a message that you are being irresponsible. A live chat would not work here. The feeling of being heard is very nice, especially if you have to defend yourself if you are being accused of something but it is actually a completely different situation.
> Have a choice between: calling, email or Zoom.

Question 18: How big is the problem if irresponsible users are wrongly marked as responsible, also known as a false negative (so they continue to use their account and the services), do you see solutions for this?

> There must be a difference between 'annoying' and 'irresponsible' behavior. Suppose you are 5 minutes late, this should be reported every time. So without consequences but with clear awareness of your actions.
> You can eliminate such false negatives by taking into account less serious factors (for example driving too fast, but not immediately 10-20 km too fast). So pay more attention to the smaller violations.
> It is essential that people who are actually irresponsible are filtered out. A few wrong decisions do not matter, even if it is a false negative. You could indicate in the app that the previous person did use the car properly and behaved irresponsibly by giving descriptions of what happened and thus resolve a potential false negative.
> It must be possible to trace which person has behaved irresponsibly by more often keeping track of the moments of use, so which user was active at what time and where.

Question 19: How big is the problem if responsible users are wrongly classified as irresponsible, i.e., a false positive, and do you see any solutions for this?

> It is only a problem if you do not have the possibility to restore the situation and clear your name. Suppose you receive a message that you are braking all the time, you must be able to explain this to the employee and solve it. If something is wrong, you should always have the chance to report it.
> In principle, you are not doing anything wrong as a responsible user with a false positive. The only thing that is important is that you have to be convincing in refuting the irresponsibility that is claimed.
> You are blocked or can receive a warning from the platform, which is an unwanted outcome. But it's about the fact that you are not actually being irresponsible, that weighs more heavily.
> You are showing good behavior by paying attention to what others are doing to catch their mistakes, you must be able to indicate and substantiate this in a conversation or email. The employee must also be able to assess such a situation correctly.
> The platform should not actually block you solely on the basis of the algorithm or a notification, only after hearing you as a user. If they find that your explanation is not enough, then unfortunately you are still labeled as irresponsible.
> A notification is useful as a first step or intermediate step, for example something like "as a platform we have seen something suspicious you would like to explain that?"
> Letting someone finish their ride should be possible, that you can use the car and not the car suddenly being turned off if you are blocked. It would be the same principle as with driving lessons, you can just finish it but explain afterwards.
> A solution would be cameras on the outside of the car to see what the environment is doing. And then check the footage if blocking is considered. Then you do not even have to talk to the user and explain yourself if that is not really necessary. Cameras on the outside that only capture the environment do not make much difference for privacy. The environment can easily be included.

Question 20: How do you perceive the trade-off between false negatives and false positives? A shared mobility platform can deal with the relationship between these two in different ways and choose to reduce false negatives or false positives. Do you see a certain balance that can be achieved here?

> It is much worse if an irresponsible driver gets a good score or ranking. This has a greater effect on other participants in traffic. Especially if everyone around you drives well and you then ruin it for the others.

> The consideration depends. The effect of blocking weighs more heavily on users who are actually blocked. If someone is unjustly not blocked (so is irresponsible, but can simply continue to use the services), that person is lucky. But it is dangerous if such a person continues to display irresponsible behavior.

> In the case that you are not blocked immediately, but it goes according to a points system for example and you lose points if you behave well, it is not as bad if you show wrong behavior and do get points.

> It is worse if a bad driver gets a good score, which entails much more effect on the rest if the they do drive well and occasionally the user does not get good reviews.

> As a consumer, you do not want to be classified as a false positive because there are all kinds of unpleasant consequences attached to it, for which you as a user have to do a lot and the platform also because you are seen as irresponsible and that involves many actions. If you filter them out in advance, you will not have the problems afterwards or at least a lot less.

> It is very annoying with regard to safety if someone gets away with being a responsible user, but is irresponsible. They are not addressed and make mistakes that are not allowed. If this is not detected, it is not included in the data as incorrect driving behavior or payment behavior, which means that historical data can provide a distorted picture of reality.

> People who make mistakes but are still labeled as responsible is wrong. More attention should be paid to this and the priority should be with the platform to remove actual irresponsible people, especially in the context of the safety of the user and other road users. On the other hand, those who are not dangerous then have to constantly explain themselves unnecessarily, appeal the decision and encounter obstacles that are not necessary in the first place. As a platform, you can lose quite a few users because you do not want to constantly challenge every decision, especially if you are behaving properly.

> Both cases are bad for the company: if you get away with irresponsible behavior, it is inconvenient for the next driver and surrounding traffic, but as a good user it is also really irritating if you constantly receive those notifications of bad behavior while you are perfectly adhering to the rules.

About Algorithm Audit

Algorithm Audit is a European knowledge platform for AI bias testing and normative AI standards. The goals of the NGO are four-fold:

Knowledge platform

Bringing together experts and knowledge to foster the collective learning process on the responsible use of algorithms, see our white papers and public standards.

Normative advice commissions

Advising on ethical issues that arise in concrete algorithmic practice through deliberative and diverse normative advice commissions, resulting in algoprudence

Technical tools

Implementing and testing technical tools to detect and mitigate bias, e.g., sociotechnical evaluation of generative AI, unsupervised bias detection and synthetic data generation.

Project work

Supporting public and private sector organisations with specific questions regarding responsible use of AI, from a not-for-profit perspective.

Structural partners of Algorithm Audit

SIDN Fund

The SIDN Fund stands for a strong internet for all. The Fund invests in bold projects with added societal value that contribute to a strong internet, strong internet users, or that focus on the internet's significance for public values and society.

European AI&Society Fund

The European AI&Society Fund supports organisations from entire Europe that shape human and society centered AI policy. The Fund is a collaboration of 14 European and American philantropic organisations.

Building *AI auditing* capacity from a *not-for-profit* perspective

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Parkstraat 22, 2514 JK The Hague

Algorithm Audit

Stichting Algorithm Audit is registered as a non-profit organisation at the Dutch Chambre of Commerce under license number 83979212